JOURNAL OF HEALTH CARE FINANCE

Spring 2018

Evaluation of the Maryland Quality-Based Reimbursement Program: Does Paying for Performance Lead to Greater Improvements in Quality than Paying for Reporting?

First Author: Christian Evensen, MS, Independent Consultant

Yi Lu, PhD, American Institutes for Research

Corresponding Author: Steven Garfinkel, PhD, American Institutes for Research

Ying Wang, PhD, Independent Consultant

Dan Sherman, PhD, American Institutes for Research

Lizhong Peng, PhD, University of West Georgia

Declaration of Conflicting Interests: The Authors declare that there is no conflict of interest.

Acknowledgements

The authors wish to thank Marilyn Moon, Ushma Patel, and Brenna Raines for their contributions to the research and the staff members of the Maryland Health Services Cost Review Commission and the Maryland Health Care Commission for facilitating access to and understanding of their data and their Quality Reimbursement and Maryland Hospital Acquired Conditions systems. This work was supported by the Robert Wood Johnson Foundation [grant number 65463].

Evaluation of the Maryland Quality-Based Reimbursement Program: Does Paying for Performance Lead to Greater Improvements in Quality than Paying for Reporting?

Abstract

Background: In value-based purchasing models for health care, sometimes called pay-for-performance (P4P), providers' compensation is linked to performance on selected quality-of-care measures. In 2008, Maryland implemented the all-payer Quality Based Reimbursement (QBR) Initiative for acute inpatient care, which linked hospital payments to performance in clinical quality measures for heart attack, heart failure, pneumonia, and surgical care infection prevention (SCIP). QBR uses a budget-neutral redistribution of DRG-based revenue from poor-performing hospitals to high-performing hospitals in the form of adjustments to future DRG payment rates based on process-of-care (POC) quality indicators measured in the previous year.

Purpose: This study evaluated the effect of the QBR program on POC quality indicators that were included in the QBR reimbursement formula and spillover effects on POC indicators that were excluded.

Methodology/Approach: Using quarterly POC quality indicators from 2007-2011 from the CMS Hospital Compare website, we estimated difference-in-differences, comparative time series econometric models of the impact of QBR on individual POC measures, condition-specific composite measures, achievement of benchmarks, and quarterly changes. We compared all 47 Maryland acute-care hospitals to 54 Pennsylvania hospitals that were not subject to a systematic P4P program during the study period. The models controlled for hospital and county characteristics, with fixed effects for quarter. The regression coefficient for the interaction of the Maryland/Pennsylvania and pre/post indicator explanatory variables indicated the effect of QBR.

Results: We found no systematic effect of QBR on POC quality indicators used in the reimbursement formula nor spillover effects on indicators that were not included in the formula. We noted from box plots of trends for each indicator that low-performing Maryland hospitals seemed to be improving more than low-performing Pennsylvania hospitals, so we used logistic and quantile regressions with additional outcome measures (likelihood of exceeding common benchmarks, quarterly change) to determine if this apparent pattern was systematically associated with QBR, but were unsuccessful. The few significant indicators of a QBR effect were more likely caused by random variation in significance than by a systematic effect.

Conclusion: Previous research on revenue-redistribution P4P programs has been mixed—about as many studies showing favorable outcomes as unfavorable outcomes—so there is no consistent evidence to support the value of these program. Our study is consistent with previous studies that found no effects, although we think that the hypothesis that low performing hospitals might benefit from further investigation.

Practice Implications: This study adds to the evidence suggesting that revenue-redistribution P4P programs based on POC measures do not produce sufficient benefit to justify their cost.

Evaluation of the Maryland Quality-Based Reimbursement Program: Does Paying for Performance Lead to Greater Improvements in Quality than Paying for Reporting?

Introduction

Historically, payers have reimbursed healthcare providers based on the quantity of the services they provided, but value-based purchasing shifts the incentives to contain cost growth and promote better quality of care. In value-based purchasing models, sometimes called pay-for-performance (P4P), providers' basic compensation and/or bonuses are linked to their performance on selected quality-of-care measures. This trend has accelerated with the passage of the Affordable Care Act (ACA), which included a number of provisions that encouraged the use of P4P in provider reimbursement.[1]

Despite the growing popularity of P4P programs, evidence of their impact on quality has been inconclusive.[2,3,4,5,6] Some studies of P4P programs for acute inpatient care have found favorable effects on (1) clinical process measures related to acute myocardial infarction (AMI), heart failure (HF), and community-acquired pneumonia (CAP); (2) access to care; (3) structural measures, such as keeping medical records current, attendance in monthly meetings, and self- directed leadership;

[1] Patient Protection and Affordable Care Act, 42 U.S.C. §§3001-3008 (2010),

[2] Li, J., Hurley, J., DeCicca, P., & Buckley, G. (2014). Physician Response to Pay-For-Performance: Evidence from a Natural Experiment. *Health Economics*, *23*(8), 962-978.

[3] Ryan, A. M., Blustein, J., & Casalino, L. P. (2012). Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Affairs*, *31*(4), 797-805.

[4] Werner, R. M., Kolstad, J. T., Stuart, E. A., & Polsky, D. (2011). The effect of pay-for-performance in hospitals: lessons for quality improvement. *Health Affairs*, *30*(4), 690-698.

[5] Mullen, K. J., Frank, R. G., & Rosenthal, M. B. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *The Rand Journal of Economics*, *41*(1), 64-91.

[6] Rosenthal, M. B., Frank, R. G., Li, Z., & Epstein, A. M. (2005). Early experience with pay-for-performance: from concept to practice. *JAMA*, *294*(14), 1788-1793.

and (4) outcome measures, such as in-hospital mortality.[7,8,9,10] Other studies, however, have documented mixed effects of the P4P programs. Ryan et al. (2015) found that the Medicare Hospital Value-Based Purchasing Program (HVBP) had little effect on hospital clinical quality and patient experience of care.[11] Ryan, Blustein and Casalino (2012) reported that HVBP's precursor, the Premier Hospital Quality Incentive Demonstration, did not achieve its intended effect on quality of care among low-performing hospitals.[12]

The Maryland QBR Program

In 2008, the Maryland Health Services Cost Review Commission (HSCRC) implemented the Quality Based Reimbursement (QBR) Initiative, the first of several P4P programs for acute inpatient care, which linked hospital payments to performance in clinical quality measures for heart attack, heart failure, pneumonia, and surgical care infection prevention (SCIP). Maryland hospitals began receiving QBR-adjusted payments on July 1, 2009 (the start of Maryland State Fiscal Year 2010), with incentive payment rates determined by performance on clinical process-of-care (POC) measures in calendar year (CY) 2008.[13] POC measures are derived from

[7] Lindenauer, P. K., Remus, D., Roman, S., Rothberg, M. B., Benjamin, E. M., Ma, A., & Bratzler, D. W. (2007). Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine*, *356*(5), 486-496.

[8] Collier, V. U. (2007). Use of Pay for Performance in a Community Hospital Private Hospitalist Group: A Preliminary Report. *Transactions of the American Clinical and Climatological Association*, *118*, 263–272.

[9] Grossbart, S. R. (2006). What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery. Medical Care Research and Review, 63(1_suppl), 29S-48S.

[10] Glickman, S. W., Ou, F. S., DeLong, E. R., Roe, M. T., Lytle, B. L., Mulgund, J., ... & Peterson, E. D. (2007). Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA*, *297*(21), 2373-2380.

[11] Ryan, A. M., Burgess, J. F., Pesko, M. F., Borden, W. B., & Dimick, J. B. (2015). The Early Effects of Medicare's Mandatory Hospital Pay-for-Performance Program. *Health services research*, *50*(1), 81-97.

[12] Ryan, A. M., Blustein, J., & Casalino, L. P. (2012). Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Affairs*, *31*(4), 797-805.

[13] Maryland Hospital Services Cost Review Commission (Maryland HSCRC). (2011). Maryland Report on Hospital Payments linked with Performance Initiatives and Hospital Value Based Purchasing Program Exemption Request Pursuant to Section 1886(o)(1)(C)(iv) of Social Security Act. URL: http: //www:hscrc:state:md:us/documents/HSCRC Initiatives/QualityImprovement/11-

administrative data; e.g., percentage of patients who present in the emergency department with chest pain who receive an electrocardiogram within 10 minutes. In Maryland FY 2012, the HSCRC added patient experience-of-care measures derived from the Consumer Assessment of Healthcare Providers and Systems Hospital Survey (HCAHPS®) to the QBR program.

Under the QBR, each Maryland acute care hospital was scored on POC and HCAHPS measures both for "attainment" (compared to the average of other hospitals) and "improvement" (compared to their own past performance). The combined score is then translated into a payment adjustment factor applied to the base reimbursement for each discharge, which is determined by its All Payer Refined Diagnosis Related Group (APR-DRG) assignment. A fixed percentage of the APR-DRG based payment is withheld from payment for each discharge at each hospital and, using a linear exchange function, is redistributed from low- to high-performing hospitals in a budget-neutral manner. The amount of payment "at risk" (i.e. withheld) was 0.5 percent of all hospital payments during 2009-11 (Maryland HSCRC, 2014) and the positive or negative adjustment was made to reimbursement rates for all inpatient discharges at participating hospitals in the subsequent fiscal year. Additional detail about the QBR can be found at the HSCRC website[14] and in previous publications.[15,16]

Because Maryland is the only state with authority to set acute inpatient care reimbursement rates for all payers including Medicare, all patients at non-Federal acute care hospitals are subject to the incentives established by the reimbursement system. Thus, the impact of the QBR program should not be diluted by excluding a substantial portion of acute inpatient discharges (i.e., those covered by Medicare) from the state incentive program and it should be more likely to have a demonstrable impact than similar programs that are limited to specified insurers. The majority of previous studies have focused on private insurers' P4P programs[17] and thus the inclusion of commercial, Medicaid and Medicare payers subject to the same incentive program is an important contribution to the growing literature on the effects of P4P programs on quality of care. Moreover, we examine the effect of the QBR on the quality of care for low- and high-performing hospitals separately and

---

11/MarylandHSCRC CostOutcomesRptVBP ExemptionRequest2011-09-30:pdf, accessed on September 15, 2014

[14] HSCRC QBR website: http://www.hscrc.state.md.us/Pages/init_qi_qbr.aspx

[15] Calikoglu, Sule, Robert Murray and Dianne Feeney, Hospital Pay-For-Performance Programs In Maryland Produced Strong Results, Including Reduced Hospital-Acquired Conditions, Health Affairs, 31, no.12 (2012):2649-2658. doi: 10.1377/hlthaff.2012.0357.

[16] Garfinkel, et al., Linking Reimbursement to Performance in Acute Care Hospitals: Lessons from Maryland's Implementation Experience. Vol. 43, No. 3, Winter 2017.

[17] For example, Li, J., Hurley, J., DeCicca, P., & Buckley, G. (2014). Physician Response to Pay-For-Performance: Evidence From A Natural Experiment. *Health Economics*, *23*(8), 962-978; Mullen, K. J., Frank, R. G., & Rosenthal, M. B. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. *The Rand Journal of Economics*, *41*(1), 64-91.

search for possible spill-over effects on POC measures that were not used to adjust reimbursement rates.

Methods

Research Questions

The QBR program is based on the assumption that by tying payment rates to quality improvement, the overall quality of care and value received by patients and other payers will improve. However, QBR might also have consequences for the quality of care that is not linked to revenue (i.e., spillover effects on excluded measures). QBR might affect excluded measures favorably if it enhanced the overall culture of quality at participating hospitals. QBR might affect excluded measures unfavorably if it caused hospitals to offset the additional attention paid to included measures by neglecting excluded measures. Thus, we studied three research questions (RQs):

1. Did the Maryland QBR program cause hospital quality measures to change more than they would have in the absence of the program?

2. How did QBR measures change?

3. Were there spillover effects on hospital quality measures that were not included in the QBR program?

Study Design

We tested the effect of the QBR program on hospital quality using a difference-in-differences approach that compared changes in quality measures for hospitals in Maryland to changes in the same measures for neighboring Pennsylvania hospitals before and after the implementation of the QBR. We chose Pennsylvania as a control state for two reasons. First, most acute care hospitals in Pennsylvania were not subject to statewide P4P programs until after our study period ended—October 1, 2012, when the Medicare HVBP program went in to effect. Therefore, comparison hospitals in Pennsylvania did not have a comparable reimbursement-based financial incentive to improve quality of care during the first three years of QBR operations in Maryland.[18] Second, because health care quality and cost vary by geographic region of the United States for reasons that are not well understood, using a contiguous state with similar population and acute care hospital characteristics provided a more rigorous comparison than comparing to the entire United States. Because quarterly data were available, we specified the difference-in-differences model as a comparative interrupted time series.

Data Sources

As part of the Hospital Quality Initiative announced by the U.S. Department of Health and Human Services (HHS) in 2001, CMS's Hospital Compare website began to publish information on process of care, patient experience, and pricing and cost of medical services at hospitals nationwide. We obtained the POC and patient experience data from the archives available on the

[18] As described below, 41 western Pennsylvania hospitals participating in a sub-state regional quality improvement program were excluded from our comparison group. Maryland remains exempt from the Medicare HVBP due to the similar designs of the QBR and the HVBP.

Hospital Compare website.[19] These data are refreshed quarterly, but in most cases the POC measures represent 12-month moving averages, instead of actual quarterly data. We discuss the implication of CMS's scoring approach for our findings in the analysis section. We obtained the county characteristics from the U.S. Census Bureau and the hospital characteristics from HSCRC and the Pennsylvania Department of Health.

Study Sample

The initial study sample included all 47 acute care hospitals in Maryland and 60 eligible acute care hospitals in Pennsylvania. Although there were 101 acute care hospitals in Pennsylvania at the time of the study, 41 hospitals in western Pennsylvania were participating in a large private P4P program so were considered ineligible for the comparison group.[20]

Of the 47 eligible hospitals in Maryland, 44 had 16 full quarters of data and one hospital had 9 full quarters of data.[21] Of the 60 eligible hospitals in Pennsylvania, 54 had 16 full quarters of data. We defined "16 full quarters" as having data for at least one of the 25 quality measures for all 16 quarters. Several POC measures were only reported by a small number of hospitals and a few patient experience measures were not available in Hospital Compare for the entire study period.[22] We included such measures and hospitals in our sample because our statistical approach does not require observations for all measures and hospitals in every quarter.

The final sample consisted of 1,577 hospital-quarters from 47 Maryland hospitals and 54 Pennsylvania hospitals, spanning the 16 calendar quarters from the third quarter of CY 2007 (beginning July 1, 2007) through the second quarter of CY 2011 (ending June 30, 2011). QBR took effect on July 1, 2009, so we had an equal number of calendar quarters (8) in the pre- and post-QBR periods.

Hospital Quality Outcome Measures

We analyzed 25 clinical POC measures endorsed by the Hospital Quality Alliance, which indicate adherence to guidelines for acute myocardial infarction (AMI, 7 measures), congestive heart

---

[19] The website can be accessed at https://www.medicare.gov/hospitalcompare/search.html

[20] Around the time that the MD QBR program was being implemented (2007 to 2010), the QualityBLUE Hospital Pay for Performance Program was beginning the third stage of a three-stage implementation in Pennsylvania. This program included 41 Hospitals in the western and central part of Pennsylvania. These hospitals were excluded from the study.

[21] The number of hospitals in Maryland decreased by one after the third quarter of 2009 as a result of the merger between Memorial Hospital and Medical Center of Cumberland. The nine quarters of data from the period before this merger were included in the study sample.

[22] For example, less than 30 hospitals reported SCIP-INF4 in any given quarter in both Maryland and Pennsylvania; SCIP-INF6 and SCIP-CARD2 are missing for all hospitals for the 6 quarters prior to the first quarter of 2009; AMI-7a is only reported by less than 20 hospitals in both Maryland and Pennsylvania in any given quarter; And SCIP_INF_4 and SCIP_INF_6 were not reported until Q1 of 2008.

failure (HF, 4 measures), pneumonia (PN, 6 measures), and the prevention of surgical infections (the surgical care improvement project, or SCIP, 8 measures). Each of these 25 POC measures indicates the proportion of "opportunities" (the denominator) where an applicable standard-of-care criterion was met ("successes," the numerator). We used hospital-level data from CMS instead of individual-level data from Maryland HSCRC, because the CMS data were available for a longer time period and we had no access to individual-level data for Pennsylvania. The long timeframe and the large number of hospitals in Maryland and Pennsylvania make Hospital Compare, which was developed for public reporting of quality, a suitable data source for research on hospital-level quality using the difference-in-differences strategy.

Not all 25 POC measures reported on Hospital Compare were included in the QBR program. Some measures were retired by the Maryland HSCRC early in the study period, while others were phased in just before the QBR program commenced in the third quarter of CY 2009 or later.[23] Thus, our evaluation of the primary effect of QBR focused on the 17 POC measures that were included in the QBR with enough data in the pre-QBR period. However, to search for secondary spillover effects (RQ3), we included the eight measures from Hospital Compare that were not used for the QBR program.

In addition to the individual POC measures, we constructed four composite condition-specific scores by averaging all POC measures within each clinical condition (AMI, HF, PN, and SCIP).[24] We also used the 12 POC measures included in the Medicare HVBP program in FFY 2013 to create an overall, cross-condition score for each hospital-quarter.[25] Finally, for each of these five summary condition scores, we created a set of four binary variables for each hospital-quarter, which indicate whether or not the hospital score in a given quarter surpassed each of four 'benchmark' scores. These benchmark scores were calculated for each hospital at baseline (the third quarter of CY 2007): two are defined as low benchmarks (one national and one state-level) and the other two are defined as high benchmarks (one national and one state-level). The four benchmarks are: (1) the national mean of all hospitals submitting data to Hospital Compare in the baseline quarter; (2) the median scores (50th percentile) among all hospitals in each state (Maryland and Pennsylvania); (3) the mean score among all hospitals in the top decile nationwide in the baseline quarter; and (4) the mean score among all hospitals in the top decile in each state in the baseline quarter. We used these binary indicators to estimate the impact of the QBR on the probability that hospitals surpassed these four benchmarks in each quarter. These binary indicators

[23] The QBR measures have 3 distinct phases: the base period, the measurement period, and the rating period. It is only during the rating period that the hospitals were being reimbursed based on their performance on these measures.

[24]When computing the condition scores, we included only QBR POC measures that were used to determine payment, so as to not confound performance on measures that were part of the payment algorithm during the study period with those that were not.

[25] These measures include AMI-7a, AMI-8a, HF-1, PN-3b, PN-6, SCIP-INF1, SCIP-INF2, SCIP-INF3, SCIP-INF4, SCIP-VTE1, SCIP-VTE2, and SCIP-CARD2. Note that not all of these items are included in the QBR program. Please refer to Table 3 for more detailed descriptions of these POC measures.

provide a more meaningful measure of improvement within the context of P4P than the individual measures, because P4P programs that tie incentive reimbursement (both rewards and penalties) to a target benchmark have been shown to result in improvements in quality.[26]

Explanatory Variables

We created two binary indicators (dummy variables) to test our hypothesis: one for state (PA = 0, MD = 1) and another to indicate if the quarter occurred during the pre- or post-QBR implementation period (pre=0, post=1). The pre-intervention period included the third quarter of CY 2007 through the second quarter of CY 2009 and the post-intervention period included the third quarter of CY 2009 through the second quarter of CY 2011. We controlled for characteristics of the hospital and the county in which the hospital is located by obtaining or constructing variables for bed size for both medical-surgical (MSG) and pediatric (PED) services,[27] patient volume, revenue, teaching status, and the percent of the county population eligible for medical assistance.[28] Finally, we constructed a set of dummy variables for each of the 15 quarters following the baseline quarter, which served as the reference category. We included these quarter dummies as fixed effects to capture the unobserved impact of the maturing culture of quality generally on POC measures during the study period in both states.

Analytic Model

We estimated a set of linear difference-in-difference, comparative interrupted time series models for the individual and composite POC measures. Equation 1 illustrates the general model:

$$Y_{jt} = \beta_0 + \beta_1 X_{jt} + \beta_2 MD_j + \beta_3 MD_j \times Post_t + Quarter\ FE_t + \epsilon_{jt} \qquad (1)$$

Where

- The unit of analysis is the hospital-quarter (i.e., each hospital's data for a single quarter) for both MD and PA hospitals
- Y_{jt} = one of the individual or composite POC measures described above for hospital *j* in each quarter *t*, defined as an integer between 0 and 100% (successes/opportunities)
- X_{jt} = a set of variables representing the hospital characteristics described above (i.e., MSG and PED bed counts, teaching status, and the percent of the county population eligible for medical assistance)
- MD_j = 0/1 state indicator of whether the hospital *j* is in PA (0) or MD (1)

[26] Mullen, K. J., Frank, R. G., & Rosenthal, M. B. (2010). Can you get what you pay for? Pay-for-performance and the quality of healthcare providers. The Rand Journal of Economics, 41(1), 64-91.

[27] We set this variable to zero for hospitals that do not have pediatric beds.

[28] Volume, revenue, and bed size were all highly correlated (> 0.90). To avoid problems with multicollinearity, we decided to only include the bed size variables in the regression models.

- $Post_t$ = 0/1 time-period indicator of whether the quarter t was before (0) or after (1) QBR began
- $Quarter\ FE_t$ = quarter fixed effect to control for unobserved factors that may impact the outcome of interest (e.g. maturation)
- βs = the regression coefficients and ε = unexplained variation.

The coefficient estimate for the interaction between state and time period (β_3) captures the effect of the QBR program under the assumption that the POC measures would have evolved similarly in both states in the absence of the QBR program. CMS's use of 12-month moving averages tends to smooth quarterly variability and correlate the POC values in each quarter with those from the quarters that surround it. To account for this serial correlation in the outcome variables, we estimated cluster-robust standard errors at the hospital level.[29]

We also conducted three robustness checks with the same specification listed in equation (1) on a range of outcome measures including (1) the individual and composite POC measures, (2) probabilities of outperforming state and national benchmarks, and (3) quarterly changes in the composite POC measures. Specifically, we used logistic regression to examine whether Maryland hospitals are more likely to exceed the national and state benchmarks (e.g. median, mean, and top 10%). Similar to Ryan et. al.,[30] we examined whether QBR affects low and high performing hospitals differently using a quantile regression model. We estimated the effect of QBR with respect to specific percentiles of the composite POC measures (i.e. 10th, 25th, 50th, 75th and 90th percentiles).

Results

Descriptive Statistics

Table 1 contains summary statistics on hospital characteristics. Although the distributions of some characteristics differ slightly, the eligible hospitals in Maryland and Pennsylvania are generally comparable. The only statistically significant difference is that 19 percent of hospitals in Pennsylvania are for-profit, compared to only 5 percent in Maryland.

Table 2 summarizes characteristics of the counties in which these hospitals operate, as well as the distribution of hospitals across counties that have been categorized based on population size and percent of population eligible for medical assistance. The eligible hospitals operate in 21 out of 24 Maryland counties (including Baltimore City) and 16 out of 67 Pennsylvania counties. Although

[29] Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50(2), 317-372.

[30] Ryan, Andrew M., James F. Burgess, Michael F. Pesko, William B. Borden, and Justin B. Dimick. "The Early Effects of Medicare's Mandatory Hospital Pay-for-Performance Program." *Health Services Research* 50, no. 1 (2015): 81-97. https://onlinelibrary.wiley.com/doi/full/10.1111/1475-6773.12206

counties do not strictly represent hospital markets, we assigned each hospital to the one county in which its principal address is located as a proxy for the socioeconomic status of the residents served by the hospital. As shown in Table 2, about one in ten Maryland hospitals operate in a relatively

Table 1. Characteristics of eligible hospitals in Maryland and Pennsylvania

Hospital Characteristics	***MD (N=45)***	***PA (N=54)***	***p-value***
Med-Surg Bed-size – Mean	150	165	0.5358
Med-Surg Bed-size – Distribution			
< 100 beds	33%	31%	0.959
100 to 199 beds	47%	46%	
200+ beds	20%	22%	
Has Pediatric Beds	51%	39%	0.2109
Pediatric Bed-size – Mean	7	5	0.4228
Pediatric Bed-size – Distribution			
< 10 beds	67%	52%	0.308
10 to 19 beds	17%	36%	
20+ beds	17%	12%	
Ownership (as of Q2 2011)			
Proprietary	5%	19%	0.036
Voluntary Nonprofit	95%	81%	
Teaching Hospital	22%	19%	0.6517

Table 2. Characteristics of counties in which eligible hospitals are located

County Characteristics	***MD***	***PA***	***p-value***
Number of Counties Represented in Study	21	16	
County Population – Mean*	264,701	380,279	0.3069
County Population - Range	20K to 964K	28K to 1.5M	
% of Residents Eligible for MA – Mean	15%	16%	0.7282
% of Residents Eligible for MA – Range	7% to 33%	7% to 33%	
Distribution of Hospitals by County Population			
< 50K	11%	2%	<0.001
50K to 499K	29%	52%	
500K to 1 Million	60%	26%	
1 Million +	0%	20%	
Distribution of Hospitals by % of County Residents Eligible for Medical Assistance			
< 10%	11%	26%	0.05
10% to 19%	56%	54%	
20% to 29%	9%	0%	
30% +	24%	20%	

* Based on annual published county-level population estimates from the U.S. Census Bureau; averaged across 4 years (FY 2008 to FY 2011)

small county (less than 50,000 in population), compared to only two percent of Pennsylvania hospitals, and none of the Maryland hospitals operate in counties with over one million residents. Also, a greater proportion of Maryland hospitals operate in counties where 20 percent or more of the residents are eligible for medical assistance.

Main Findings

RQ1. Did the Maryland QBR program cause hospital quality measures to change more than they would have in the absence of the program? Table 3 displays the results from the difference-in-differences models for the 17 individual POC measures that are included in the QBR. These models compare the average effects over all Maryland hospitals to all eligible Pennsylvania hospitals. Relative to Pennsylvania hospitals, Maryland hospitals differed significantly ($p \leq 0.05$) on only three of the 17 POC measures. Of those three measures, Maryland hospitals performed worse than Pennsylvania hospitals on two: HF-3 and SCIP-INF1. Maryland hospitals performed better on influenza vaccination rates (PN-7).

Table 3. Difference-in-differences estimates of the effect of the QBR program on POC measures included in the QBR reimbursement formula

Measures Included in QBR	***Estimated effect***
AMI-1: Aspirin at Arrival (N=1,545)	-0.17 (0.96)
AMI-2: AMI-2: Aspirin Prescribed at Discharge (N=1,538)	-0.65 (1.20)
AMI-3: ACEI or ARB for LVSD (N=1,483)	0.09 (1.99)
AMI-4: Adult Smoking Cessation Advice/Counseling (N=1,457)	0.49 (1.60)
AMI-5: Beta Blocker Prescribed at Discharge (N=1, 539)	-1.02 (1.32)
HF-1: Discharge Instructions (N=1,545)	-3.30 (1.94)
HF-2: Evaluation of LVS Function (N=1,545)	-0.19 (0.77)
HF-3: ACEI or ARB for LVSD (N=1,545)	-3.06 (1.00)**
HF-4: Adult Smoking Cessation Advice/Counseling (N=1,545)	-1.66 (1.48)
PN_2: Pneumococcal Vaccination (N=1,545)	2.01 (1.57)
PN_3b: ER Blood Culture Prior to Antibiotic Received in Hospital (N=1,542)	-0.36 (0.80)
PN_4: Adult Smoking Cessation Advice/Counseling (N=1,544)	-0.99 (1.28)
PN_5c: Initial Antibiotic(s) within 6 Hours After Arrival (N=1,527)	0.28 (0.62)
PN_7: Given Influenza Vaccine (N=1,545)	4.48 (1.64)**
SCIP-INF1: Antibiotic rcvd within 1 Hour b/4 Surgical Incision (N=1,572)	-2.88 (1.10)*
SCIP-INF2: Antibiotic - Right Selection (N=1,572)	0.47 (0.63)
SCIP-INF3: Antibiotics Discontinued within 24 hrs After Surgery (1,572)	-0.45 (1.74)

Note: the number of observations varies from 234 to 1,577 hospital-quarters across outcomes due to missing values. All regression models control for characteristics of the hospital and the county in which the hospital is located including bed size for both medical-surgical (MSG) and pediatric (PED) services, teaching status, and the percent of the county population eligible for medical assistance. We also include quarter fixed effects. Levels of significance: **p<0.01, *p<0.05. Cluster-robust standard errors are reported in parentheses next to Estimated Effect.

RQ2. How did QBR measures change? To shed additional light on these average effects, we examined box plots of the quarterly trends in the outcomes separately for the states. Because the distributions of the POC measures are skewed, we focused on the trend in median scores from quarter to quarter. For many measures, the trend over time was erratic or flat. However, the distributions of 10 of the 25 POC measures (both included and excluded in QBR), especially the HF and SCIP measures, narrowed in the post-period for Maryland hospitals compared to Pennsylvania hospitals. They narrowed because improvement in the scores of the lower performing hospitals in the post-intervention period was more pronounced in Maryland than in Pennsylvania. That is, hospitals with relatively low scores seemed to improve more in Maryland under QBR than in Pennsylvania without QBR.

Figure 1 illustrates this pattern for one QBR measure: SCIP-INF3, Antibiotics Discontinued within 24 Hours after Surgery. The scores started high in both states, exhibiting a slight upward trend in the median and mean. There are a number of outlier hospitals in the beginning of the pre-intervention period, though they start to shift toward the mean by the end of the pre-intervention period in both states. In the post-intervention period, the means and medians continue to improve at the same rate in both states, but hospitals at the bottom of the distribution in Maryland continued to move up, decreasing the overall variance, while performance of hospitals at the bottom of the Pennsylvania distribution started to decline, causing the overall variance to increase toward pre-intervention levels. By the end of the post-intervention period, the bottom end of the distribution in Pennsylvania is well below that of Maryland.[31] We observed a similar pattern with seven other QBR measures[32] and two non-QBR measures.[33] Based on the changes in these distributions, we hypothesized that the QBR and similar P4P programs based on POC measures might be most likely to influence low to medium performers. Our initial focus on the means as the outcome of interest might have masked an effect because of the outliers. Thus, we estimated three additional sets of statistical models to determine if this hypothesis, based on our anecdotal observation of the box plots, could be confirmed as a systematic effect over all hospitals and outcomes.

All three models use the same set of explanatory variables we used in Equation 1, but different outcome variables and the appropriate form (linear, logistic or quantile regression) depending on the structure of the outcome measures.

First, we evaluated the effect of QBR on the probability that hospital scores exceeded national and state benchmarks, using logistic regression models. Attaining or exceeding state, regional, or national benchmarks is a typical criterion used to assign bonus payments in P4P systems. We tested

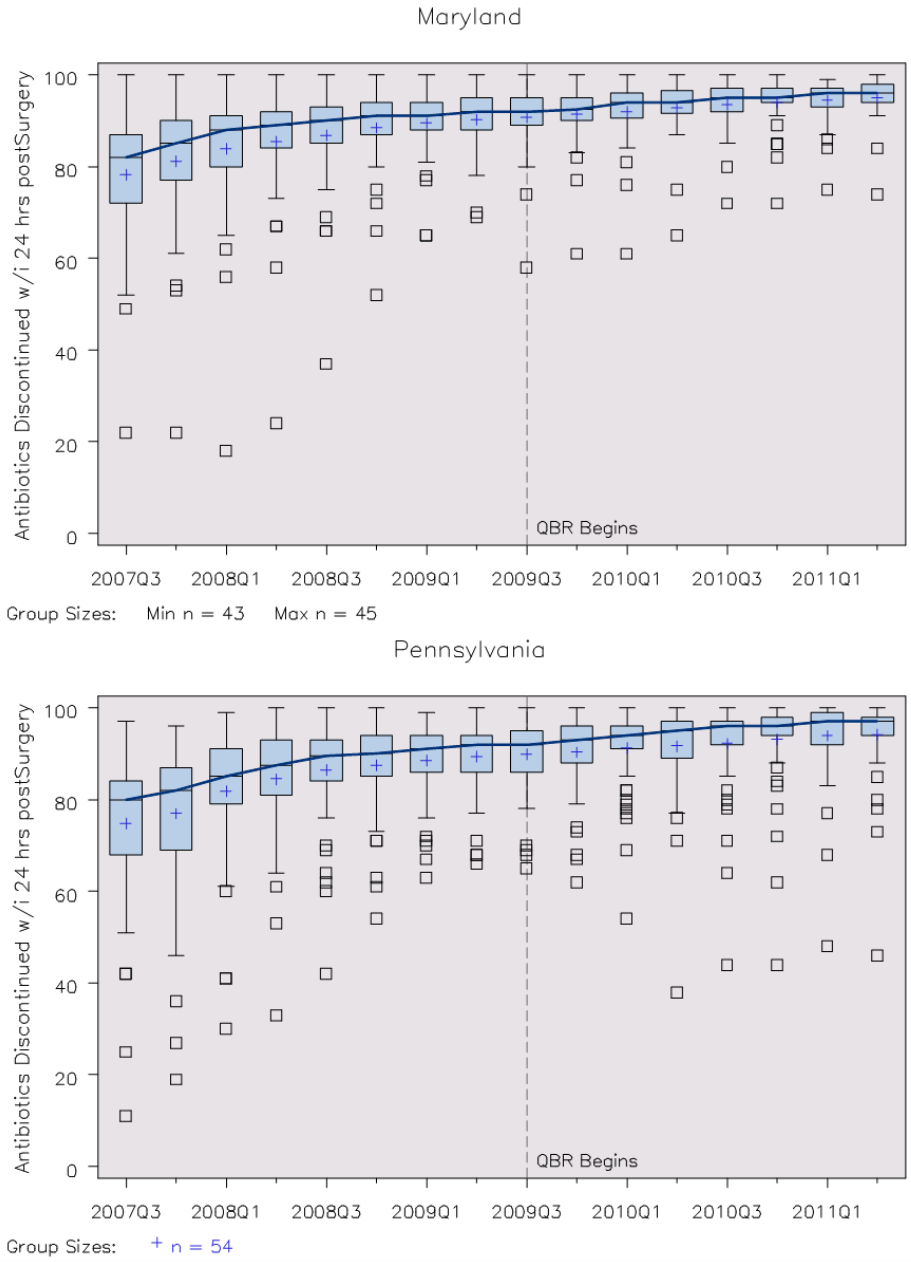
[31] The minimum score in Maryland was 74, compared to 46 for Pennsylvania; the median in Maryland was 91, but only 78 for Pennsylvania. The two states have a similar distribution above the 25th percentile.

[32] The eight QBR POC measures that exhibited this pattern include AMI4 (smoking cessation counseling), HF1 (discharge instructions), HF2 (evaluation of LVS function), HF4 (smoking cessation counseling), PN3b (ER blood culture prior to giving antibiotic), PN4 (smoking cessation counseling), PN7 (flu vaccine), SCIP_INF1 (given antibiotic 1 hour b/4 surgery).

[33] SCIP_INF6 (safer hair removal) and SCIP_CARD2 (received beta blocker in perioperative period).

QBR with state and national benchmarks at low and high levels. The low benchmarks included the national means and state-specific medians for Maryland and Pennsylvania. QBR uses the statewide mean as its definition of attainment. The high benchmarks included the 90th percentiles of the national and state-specific distributions. We chose a high bar under the assumption that, as P4P programs mature, hospitals will have to meet increasingly stringent standards to justify rewards. We used the four condition-specific composite measures as our outcome variables in these models, because composite scores tend to be more stable and reliable than individual measures. They were computed as the mean of both included and excluded

Figure 1. Change in distribution in QBR measure SCIP-Inf3 by state



individual measures. Combining included and excluded measures in the composite is a way to determine if QBR has created a culture of quality that transcends the individual measures included in the P4P determination.

From Table 4, we see that, after the introduction of the QBR program, hospitals in Maryland were no more likely than hospitals in Pennsylvania to surpass national mean or state median benchmarks in any of the four condition areas. QBR affected only the probability of surpassing the 90th percentile of their respective state distributions for AMI and HF scores, but Maryland hospitals were less likely to do so than Pennsylvania hospitals.

Table 4. Difference-in-differences estimates of the effect of the QBR program on the probability of exceeding quality-of-care benchmarks by clinical condition

Outcome	Low Benchmarks		High Benchmarks	
	Above national average	Above state's median	National top 10%	State's top 10%
AMI	0.41 (1.29)	0.08 (0.39)	-0.39 (0.30)	-1.08** (0.39)
HF	N/A	0.64 (0.49)	N/A	-0.67 (0.35)
PN	0.46 (0.64)	0.42 (0.49)	N/A	0.27 (0.43)
SCIP	0.70 (1.02)	0.38 (0.74)	0.20 (0.57)	0.51 (0.47)

Notes: All regressions control for bed size for both medical-surgical (MSG) and pediatric (PED) services, patient volume, revenue, teaching status, and the percent of the county population eligible for medical assistance. We also include quarter fixed effects. Levels of significance: **p<0.01, *p<0.05. Cluster-robust standard errors are reported in parentheses. N/A denotes situations where there is not enough variation in the dependent variable, and as a result the logistic regression failed to converge. For example, among the 1545 observations for the HF National top 10% indicator, only 80 hospital-quarters surpassed the baseline national average in the post-QBR period.

After failing to find that QBR affected the probability of attaining specified low and high performance thresholds, we estimated a parallel set of models to see if QBR was associated with improvement at any place along the distribution; a lower bar than attaining or exceeding benchmarks, but one that should be more sensitive to improvement among low-performing hospitals. The ordinary least squares (OLS) model can be inefficient when there are many outliers and a highly non-normal error term. Therefore, we applied a quantile regression model using the same explanatory and outcome variables to estimate the impact of QBR on subsets of the distribution of performance scores using five quantiles defined by ceilings set at the 10th, 25th, 50th, 75th, and 90th percentiles. The objective is to detect if the QBR effect varied over performance-measure quantiles and if the magnitude of the effects is significantly different from that of the OLS estimates.

The regression parameters in Table 5 demonstrate the effects of QBR on the condition-specific composite performance scores based on the OLS models and the models for each of the five quantiles. Of the 24 coefficients reported in Table 5, only five are significant, with Maryland hospitals performing better than Pennsylvania hospitals in only the pneumonia composite. All the

significant coefficients are in the median and 75th percentile quantiles, suggesting impacts are occurring among Maryland hospitals in the middle to high quantiles of the performance scores. We plotted the coefficients (green solid line) in Appendix Figure 1 for the four composite outcome measures to visually illustrate how the QBR effect varies over score quantiles and how they are contrasted with the mean estimates from the OLS model (black dashed line).

Finally, we tested whether there is systematic evidence that QBR is associated with the rate of change in performance by modeling the quarter-to-quarter change in the clinical measures as our dependent variable in the linear regression model. Table 6 reports the findings. All but one of the coefficients are insignificant, but the consistently positive signs indicate a more pronounced upward trend in the performance measures under the QBR system.

Table 5. Quantile regression estimates of the effect of the QBR program on POC composite measures

	(1) OLS	(2) Q (10th percentile)	(3) Q (25th percentile)	(4) Q (50th percentile)	(5) Q (75th percentile)	(6) Q (90th percentile)
AMI	-0.370	0.577	1.304	-0.523	-0.563*	-0.209
	(0.863)	(1.905)	(1.146)	(0.321)	(0.265)	(0.150)
HF	-2.052*	-1.430	-1.575	-1.252*	-0.869*	-0.341
	(0.963)	(2.894)	(0.880)	(0.614)	(0.389)	(0.383)
PN	0.957	-1.259	0.239	1.236*	1.058*	0.517
	(0.748)	(2.258)	(0.950)	(0.538)	(0.436)	(0.405)
SCIP	-0.953	-0.383	-0.622	-0.349	-0.373	-0.166
	(0.814)	(2.254)	(0.922)	(0.351)	(0.295)	(0.423)

Notes: OLS estimates are reported in Column (1). Columns (2)-(6) report estimates from quantile regressions with selected quantiles at 10th, 25th, 50th, 75th and 90th percentiles. All regressions control for bed size for both medical-surgical (MSG) and pediatric (PED) services, teaching status, and the percent of the county population eligible for medical assistance. We also include quarter fixed effects. Levels of significance: **p<0.01, *p<0.05. Standard errors are reported in parentheses.

Table 6. Difference-in-differences estimates of the effect of the QBR program on Quarterly Changes in POC Composite measures

	(1) AMI	(2) HF	(3) PN	(4) SCIP
MD QBR Effect	0.178	0.332	0.0721	0.500*
	(0.342)	(0.253)	(0.283)	(0.213)
Observations	1448	1448	1448	1472

All regression models control for characteristics of the hospital and the county in which the hospital is located including bed size for both medical-surgical (MSG) and pediatric (PED) services, teaching status, and the percent of the county population eligible for medical assistance. We also include quarter fixed effects. Levels of significance: **p<0.01, *p<0.05. Cluster-robust standard errors are reported in parentheses.

RQ3. Were there spillover effects on hospital quality measures that were not included in the QBR program? To detect spillover effects of QBR, we looked at its impact on four kinds of outcome measures. First, we modeled QBR impact on the eight individual POC measures in our data that were excluded from the QBR payment algorithm (Table 7). By focusing only on measures that were not used to determine payment, we have a pure test of spillover effects. QBR affected only one measure—AMI-7a: Fibrinolytic Therapy within 30 Minutes of Arrival—but the effect is favorable and very large. Nevertheless, due to missing data, we have only 234 hospital-quarter observations for AMI-7a. Given the small sample and the unusual size of the estimated effect, this finding is suspect.

Although there is little evidence for spillover effects on individual POC measures, the ultimate goal of P4P programs like QBR is to create a culture of quality and improve performance across the board, not only among the measures included in the algorithm. Therefore, we used the same model to look for effects of QBR on the four condition-specific composite scores that combine included and excluded measures. In the last section of Table 7, we see that QBR had essentially no effect on these condition-specific composites.[34] QBR had an unfavorable impact on the HF composite measure at $p < 0.05$. However, because all four individual indicators that comprise the HF composite are included in the QBR, this finding reflects the unfavorable impact of QBR on included measure HF-3, previously discussed under RQ1.

Discussion

We studied the impact of the Maryland QBR program on one aspect of the quality of inpatient care: clinical process of care (POC) measures. Although we found a few significant coefficients, we found no systematic effect of QBR on quality of care measured over all eligible hospitals, which is consistent with several previous studies on the effect of P4P programs.[35,36,37] We also found no spillover effects on excluded measures, which is not surprising given there were no systematic effects on included measures.

---

[34] QBR had an unfavorable impact on the HF composite measure at $p < 0.05$. However, because all four individual indicators that comprise the HF composite are included in the QBR, this finding reflects the unfavorable impact of QBR on included measure HF-3, previously discussed under RQ1.

[35] Glickman, S. W., Ou, F. S., DeLong, E. R., Roe, M. T., Lytle, B. L., Mulgund, J., ... & Peterson, E. D. (2007). Pay for performance, quality of care, and outcomes in acute myocardial infarction. *JAMA, 297*(21), 2373-2380.

[36] Ryan, A. M., Blustein, J., & Casalino, L. P. (2012). Medicare's flagship test of pay-for-performance did not spur more rapid quality improvement among low-performing hospitals. *Health Affairs, 31*(4), 797-805.

[37] Ryan, A. M., Burgess, J. F., Pesko, M. F., Borden, W. B., & Dimick, J. B. (2015). The Early Effects of Medicare's Mandatory Hospital Pay-for-Performance Program. *Health services research, 50*(1), 81-97.

Table 7. Difference-in-differences estimates of the effect of the QBR program on POC measures excluded from QBR and composites of included and excluded measures

Measures Excluded from QBR	***Estimated Effect***
AMI-7a: Fibrinolytic Therapy within 30 Minutes of Arrival (N=234)	43.99 (13.98)**
AMI-8a: PCI within 90 Minutes of Arrival (N=853)	1.89 (2.88)
PN_6: Given the Most Appropriate Initial Antibiotic(s) (N=1,545)	-0.78 (1.02)
SCIP-INF4: Cardiac Surgery patients w/Controlled 6 A.M. Postop BG (N=613)	4.17 (7.65)
SCIP-INF6: Surgery patients with App. Hair Removal (N=1,375)	0.73 (0.81)
SCIP-VTE1: Surgery patients w/Recommended VTE Prophylaxis Ordered (N=1,509)	-0.34 (1.56)
SCIP-VTE2: Patients Received Appropriate VTE b/4 & Post-surgery (N=1,509)	-0.48 (1.75)
SCIP-CARD2: Received beta-blocker in Perioperative Period (N=952)	1.90 (1.57)
Composite Condition Scores Combining QBR and Non-QBR Measures	***Estimated Effect***
AMI condition score (N=1,545)	-0.37 (0.86)
HF condition score (N=1,545)	-2.05 (0.96)*
PN condition score (N=1,545)	0.96 (0.75)
SCIP condition score (N=1,572)	-0.95 (0.81)

Note: the number of observations varies from 234 to 1,577 hospital-quarters across outcomes due to missing values. All regression models control for characteristics of the hospital and the county in which the hospital is located including bed size for both medical-surgical (MSG) and pediatric (PED) services, teaching status, and the percent of the county population eligible for medical assistance. We also include quarter fixed effects. Levels of significance: **p<0.01, *p<0.05. Cluster-robust standard errors are reported in parentheses next to Estimated Effect.

Nonetheless, the trends in some performance measures observed in their box plots (e.g., Figure 1) suggested that QBR might be motivating lower performing hospitals to improve more quickly and to retain that improvement longer in Maryland than in comparison hospitals, which would be a favorable impact even in the absence of significant average effects. We estimated three statistical models to determine if this hypothesis could be confirmed generally. We were unable to find a systematic effect of QBR using any of the three statistical approaches. QBR hospitals were not more likely than comparison hospitals to exceed low or high benchmarks that might be used to determine bonus payments. We found limited evidence suggesting the QBR effect varies across different quantiles of the POC measures. QBR hospitals were not more likely than comparison hospitals to exhibit larger period-to-period percentage improvement.

Although we could not confirm the hypothesis that QBR was responsible for moving low performing hospitals toward the median and reducing the variability of performance, we observed this beneficial pattern in 10 of the 25 POC measures in our data so, rather than reject this hypothesis, we think it remains an intriguing pattern worthy of further study by other investigators using other data. Future evaluations of P4P programs should go beyond the search for significant average effects by investigating effects on the distribution and variance of measures included in and excluded from the reimbursement algorithm.

On the surface, this result appears to differ from one reported by HSCRC staff in 2012,[38] but the findings from both studies are actually consistent. HSCRC reported that all QBR measures improved in Maryland between 2007 and 2010, but only half improved more than in the U.S. as a whole. Only three of these national comparisons were statistically significant and for only one (flu vaccination for patients with pneumonia) did Maryland improve more than the nation. The two studies differ in several ways, which could account for differing results. HSCRC tested differences in rates of change between Maryland and U.S. average scores without controlling for the hospital and geographic characteristics included in our regression models and the focal periods differed (2007-10 vs. 2008-12). Nevertheless, the small number of inconsistent significant effects for QBR measures found in both studies lends credence to the notion that we are seeing random significance over multiple tests rather than systematic impact, regardless of method. The absence of systematic effects is also consistent with information provided by hospital managers in key informant interviews at four Maryland hospitals, published elsewhere.[39] Managers expected that QBR-type POC measures would have little impact on their operations or performance, because they were already reporting such measures to CMS under pay-for- reporting, the percentage of revenue at risk was too small, and the pay-off (adjustment to APR-DRG payment in the subsequent year) was too remote. However, HSCRC's observation that the variances of POC measures decreased over time in Maryland is consistent with the hypothesis about potential impact on low-performing hospitals and supports our recommendation for additional research.

Limitations

The reader should consider three limitations of this study. First, although we have 16 quarters of data (8 pre- and 8 post-intervention) for each hospital, the values of the POC measures for each quarter are 12-month moving averages, which tends to smooth out quarterly variation and deviation from prior trends. Although we adjusted standard errors to correct for the serial correlation in these measures, smoothing makes it less likely that small and moderate effects will be statistically significant. Thus, our analysis is necessarily more conservative than it would have been if we had values that were unique to each quarter.

Second, our findings could be sensitive to the choice of Pennsylvania hospitals as our comparison group. We considered alternatives, including hospitals in Virginia and all other states combined, but settled on Pennsylvania for two main reasons. We preferred a single comparison state, because it was feasible to identify and exclude hospitals participating in other state-wide quality improvement initiatives. Once we decided to use a single state, we searched for one that was similar to Maryland in regional location, demographic characteristics, and hospital characteristics that did not have a statewide QBR-like initiative in place during our study period, to control as well as possible for competing explanations. Pennsylvania best met these criteria after we excluded

[38] Sule Calikoglu, Robert Murray and Dianne Feeney, Hospital Pay-For-Performance Programs In Maryland Produced Strong Results, Including Reduced Hospital-Acquired Conditions, Health Affairs, 31, no.12 (2012):2649-2658. doi: 10.1377/hlthaff.2012.0357.

[39] Garfinkel, S. et al., Linking Reimbursement to Performance in Acute Care Hospitals: Lessons from Maryland's Implementation Experience. Vol. 43, No. 3, Winter 2017.

the 41 central and western Pennsylvania hospitals participating in a common quality improvement initiative. We used both time invariant and variant hospital characteristics in our models to control for residual variation in hospital-specific initiatives, which were impossible to identify individually in Pennsylvania or Maryland. Nevertheless, the choice of Pennsylvania hospitals could have affected our findings.

Third, during the period of our study, Maryland operated a companion quality improvement initiative, called the Maryland Hospital Acquired Conditions (MHAC) program, in which the same hospitals participated. MHAC used a similar reimbursement redistribution approach based on rates of preventable complications, which focus more on outcomes than process. Because their participating hospitals and time frame are identical, QBR and MHAC are confounded. Nevertheless, it is hard to imagine that the MHAC program could have affected QBR measures more than the QBR program itself. Thus, we do not believe that the absence of QBR effects is the result of MHAC.

Conclusion

Despite the popularity and conceptual appeal of P4P programs using revenue redistribution based on POC measures, we found no conclusive evidence that Maryland's QBR program affected quality of care. Nor were we able to confirm the hypothesis suggested by observation of the data for some measures that QBR motivated low performers to improve. Thus, it appears that the QBR approach provided an insufficient incentive to improve quality.[40] Nevertheless, payers are accelerating the integration of QBR-type P4P elements in their payment systems, so reimbursement is increasingly affected by POC quality measures. For example, the payment "at risk" for the Medicare HVBP increased from 0.5 percent to 2 percent between Federal FY 13 and 17. Given the rising stakes presented by the rapid expansion of P4P reimbursement, it has become increasingly important to understand the effectiveness of such programs in improving quality of care. If P4P programs are not effective, they might actually reduce social welfare, because the cost incurred by providers to participate in these programs, including both cash and opportunity costs, may not be offset by adequate benefits. Future research should compare alternative types of P4P programs; examine their impact on other aspects of care, such as patient outcome and patient experience measures; and go beyond average treatment effects to study their effects on the lowest performers.

[40] Maryland continues to operate the QBR and MHAC programs, but added a global budget demonstration in 2014, which theoretically provides a stronger incentive. The global budget model is currently being evaluated.

Additional Information About the Authors

First Author: Christian Evensen, MS, Independent Consultant
ctevensen.nc@gmail.com, (919) 272-0505

Yi Lu, PhD, American Institutes for Research
ylu@air.org, (202) 403-6807

Corresponding Author: Steven Garfinkel, PhD, American Institutes for Research
100 Europa Drive, Suite 315, Chapel Hill, NC 27517
sgarfinkel@air.org, 919-918-2306

Ying Wang, PhD, Independent Consultant
flwying@gmail.com, (443) 562-0051

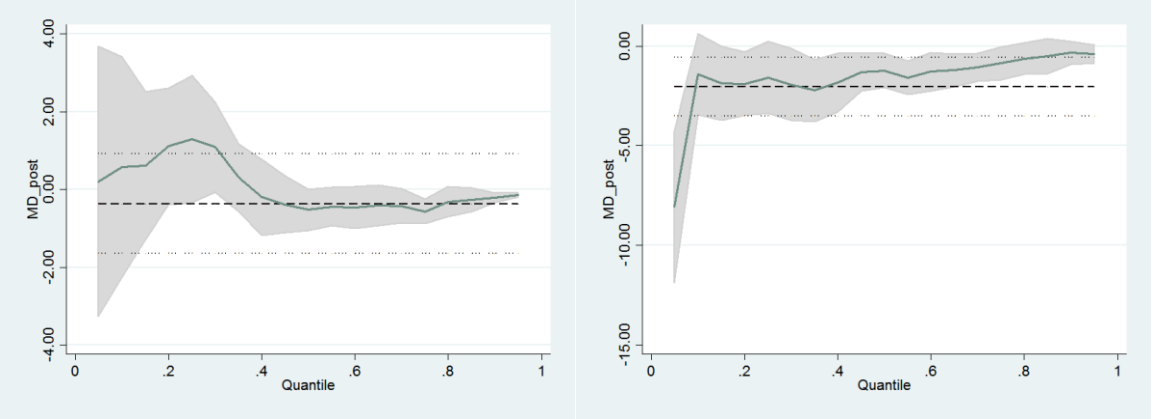
Dan Sherman, PhD, American Institutes for Research
dsherman@air.org, (202) 403-5401

Lizhong Peng, PhD, University of West Georgia
lpeng@westga.edu, (678) 839-2498

Declaration of Conflicting Interests: The Authors declare that there is no conflict of interest.

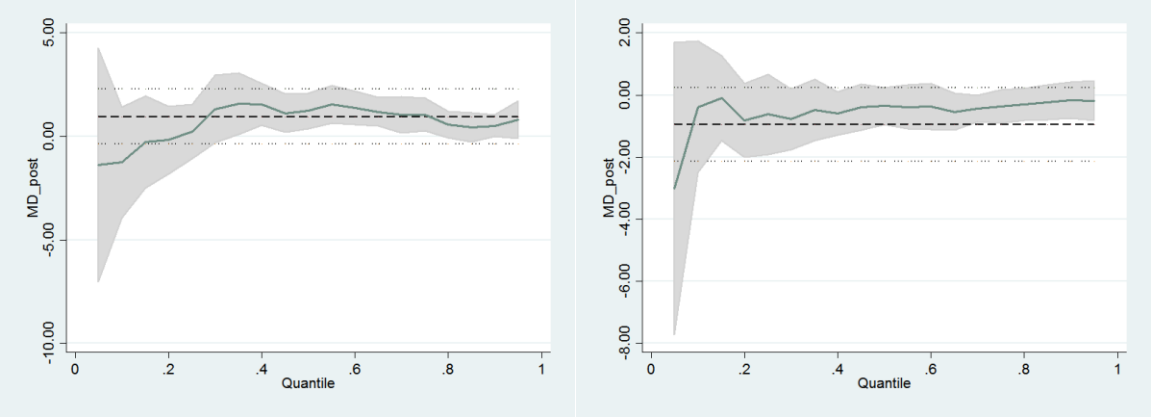
Appendix. Figure

Quantile regression estimates of the effect of the QBR program on condition specific composite measures (using 5 quantiles to define the distribution)



Composite AMI condition score (N=1,545)

Composite HF condition score (N=1,545)



Composite PN condition score (N=1,545)

Composite SCIP condition score (N=1,572)