Real-time Feedback in Pay-for-Performance: Does More Information Lead to Improvement?

Amelia M. Bond, PhD[1,2], Kevin G. Volpp, MD PhD[3,4,5], Ezekiel J. Emanuel, MD PhD[3,4,5], Kristen Caldarella, MHA[5], Amanda Hodlofski, MPH[6], Lee Sacks, MD[7], Pankaj Patel, MD MSc[7], Kara Sokol, MHSA/MPP[7], Salvatore Vittore, CPA[7], Don Calgano, MBA[7], Carrie Nelson, MD[7], Kevin Weng, MS[7], Andrea Troxel, ScD[8], and Amol Navathe, MD PhD[3,4,5]

[1]Health Care Management, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA; [2]Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA; [3]Leonard Davis Institute of Health Economics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA; [4]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA; [5]Division of Health Policy, University of Pennsylvania, Philadelphia, PA, USA; [6]HealthCore, Inc., Wilmington, DE, USA; [7]Advocate Health System, Chicago, IL, USA; [8]Department of Population Health, New York University School of Medicine, New York, NY, USA.

BACKGROUND: Pay-for-performance (P4P) has been used expansively to improve quality of care delivered by physicians. However, to what extent P4P works through the provision of information versus financial incentives is poorly understood.

OBJECTIVE: To determine whether an increase in information feedback without changes to financial incentives resulted in improved physician performance within an existing P4P program.

INTERVENTION/EXPOSURE: Implementation of a new registry enabling real-time feedback to physicians on quality measure performance.

DESIGN: Observational, predictive piecewise model at the physician-measure level to examine whether registry introduction associated with performance changes. We used detailed physician quality measure data 3 years prior to registry implementation (2010–2012) and 2 years after implementation (2014–2015). We also linked physician-level data including age, gender, and board certification; group-level data including registry click rates; and patient panel data including chronic conditions.

PARTICIPANTS: Four hundred thirty-four physicians continuously affiliated with Advocate from 2010 to 2015.

MAIN MEASURES: Physician performance on ten quality metrics.

KEY RESULTS: We found no consistent pattern of improvement associated with the availability of real-time information across ten measures. Relative to predicted performance without the registry, average performance increased for two measures (childhood immunization status—rotavirus ($p < 0.001$) and diabetes care—medical attention for nephropathy ($p = 0.024$)) and decreased for three measures (childhood immunization status—influenza ($p < 0.001$) and diabetes care—HbA1c testing ($p < 0.001$) and poor HbA1c control ($p < 0.001$)). Results were consistent for subgroup analysis on those most able to improve, i.e., physicians in the bottom tertile of performance prior to registry introduction. Physicians who improved most were in groups that accessed the registry more than those who improved least (8.0 vs 10.0 times per week, $p = 0.010$).

CONCLUSIONS: More frequent provision of information, provided in real-time, was insufficient to improve physician performance in an existing P4P program with high baseline performance. Results suggest that electronic registries may not themselves drive performance improvement. Future work should consider testing information feedback enhancements with financial incentives.

KEY WORDS: performance measurement; health information technology; physician behavior; evaluation.

J Gen Intern Med 34(9):1737–43
DOI: 10.1007/s11606-019-05004-8

INTRODUCTION

Pay-for-performance (P4P) has been used by commercial and public payers in the USA and in other nations in attempts to improve physician service quality for over a decade.[1–4] Recent US policy changes will expand P4P in Medicare through the Medicare Access and CHIP Reauthorization Act (MACRA) and its Merit Incentive Payment System (MIPS) program, one of Medicare's largest changes to physician payment in its history.[5] Yet, prior evaluations of P4P have largely focused on results with limited examination of mechanisms.[6]

Conceptually, insights from classical and behavioral economics indicate that individuals respond to extrinsic (e.g., monetary) and/or intrinsic rewards. P4P uses both types of incentives with the introduction of information on performance (intrinsic) and monetary incentives tied to this performance (extrinsic). Decomposing the effects of P4P incentives

This work was presented at the Society of General Internal Medicine in May 2017.

Electronic supplementary material *The online version of this article (https://doi.org/10.1007/s11606-019-05004-8) contains supplementary material, which is available to authorized users.*

Received June 25, 2018
Revised November 8, 2018
Accepted March 14, 2019
Published online April 30, 2019

by these mechanisms could inform future designs of P4P incentives, perhaps in ways that could enhance their impact without increasing their cost.

In this study, we examined how physicians responded to changes in intrinsic incentives (information) while holding extrinsic incentives (monetary rewards) constant. On January 1, 2014, Advocate implemented a Cerner registry moving from quarterly quality reports on paper to allowing physician real-time access to their quality scores. The P4P monetary rewards remained fixed. This "natural experiment" enables the study of the information feedback mechanism in P4P.

We estimated the overall impact of greater access to information on physician quality performance. Additionally, we estimated the impact on physicians who had the largest ability to improve following earlier literature.[7] Focusing on previously low performing physicians is particularly important in the context of Advocate Health System. Relative to other health plan and health system comparators, its scores are well above state and national averages (Fig. 1) and thus it may be more difficult for already high performing physicians to improve. Finally, we compared physician characteristics between physicians who improved the least and those who improved the most to determine characteristics associated with improved performance.

METHODS

Setting

Advocate Health System is the largest health system in Illinois, with 12 hospitals and an affiliated physician organization, Advocate Physician Partners (APP). APP is a clinically integrated network of over 4800 physicians, the vast majority of whom are affiliated and not directly employed. A bonus program for the affiliated physicians began over a decade ago, while a smaller bonus program for employed physicians began in 2012. Affiliated physicians may receive an additional 10% to 50% of their base pay through Advocate's bonus or Clinical Integration (CI) program, representing large bonuses relative to other P4P programs particularly because Advocate patients do not comprise the entire panel of affiliated physicians.[8] The CI program uses a variety of performance metrics, many of which are based on the Healthcare Effectiveness Data and Information Set (HEDIS), including registry type measures such as diabetic screening; utilization measures such as percent of generic prescription and average length of stay at the group level; and selected patient satisfaction survey measures.

APP also ran annual quality improvement (QI) programs, which primarily focused on education and outreach around the registry systems.[9] From 2010 through 2012, physicians could also participate in a full day and two half-day trainings or "collaboratives." The topics were different each year and included asthma, coronary artery disease, diabetes, and access to care. No other QI program existed 2013 onward outside of assisting practices adapt to the new electronic registry system. APP was unaware of other quality improvement programs within individual affiliated practices.[10]

An important feature of Advocate is its historic performance and longstanding exposure of its physicians to P4P incentives. Advocate collects a wide range of quality metrics, far beyond those typically collected by many health care organizations, so direct comparison of Advocate physicians to other organizations across all Advocate metrics is not possible. For the set of comparable HEDIS measures used in this study, Advocate physicians consistently score significantly higher relative to state and national averages. Advocate's process quality measure average was consistently over five percentage points

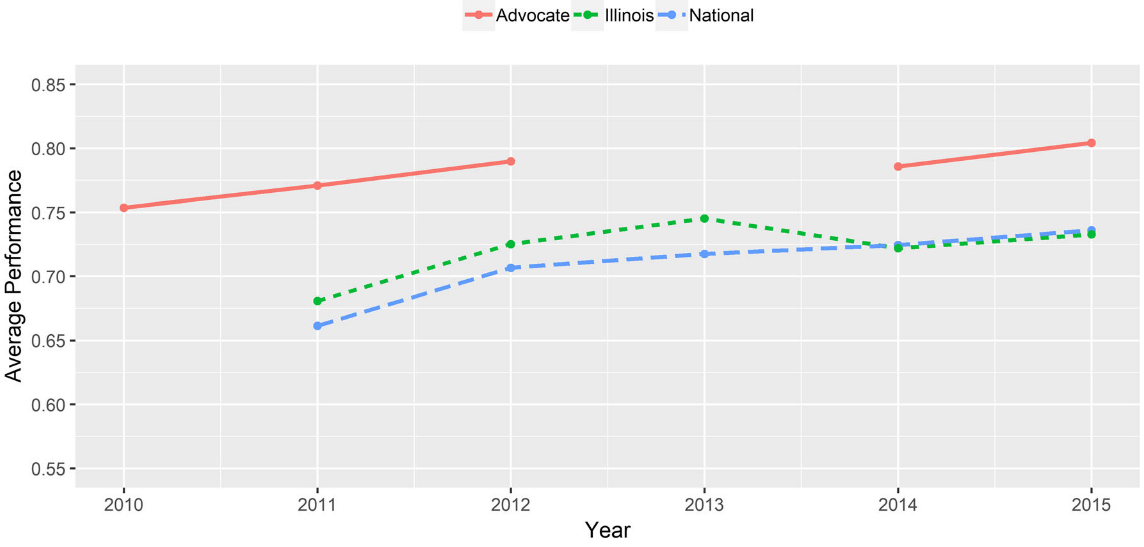

Figure 1 HEDIS performance over time. Average process quality performance includes three HEDIS diabetes care measures (HbA1c testing, eye exams, and medical attention for nephropathy) and three HEDIS childhood immunization status measures (rotavirus, influenza vaccinations, and combo 3, which includes the following immunizations DTAP, IPV, MMR, HIB, HEPATITIS B, VZV, and PCV). National- and state-level performance is the mean of all commercial (any line of business) insurance plans. National- and state-level performance is missing in 2010 because not all HEDIS quality measures were collected in 2010.

higher than the commercial HEDIS Illinois and national US averages (Fig. 1). Advocate also performed higher than state and national averages on the intermediary outcome measure diabetes care—HbA1c control. In 2012, for example, 70% of Advocate diabetic patients had HbA1c readings under 8 whereas the national and state commercial insurance averages were both 10 percentage points below.

On January 1, 2014, Advocate implemented a new Cerner registry allowing physicians' real-time access to their quality scores, seeking a way to increase quality performance across its network. Physicians and designated staff in each medical group were able to access the registry portal and visualize physician-level real-time performance on composite scores, individual measures, and detailed patient information on care gaps (see Figure A1 in the online appendix for a screen shot of the registry). Prior to electronic registry implementation, physicians received individual quality reports on paper once a quarter that simply listed each measure's score with no patient-level information.

Data

We used detailed quality-measure level data by physician from 3 years prior to the registry implementation (2010–2012) and 2 years after implementation (2014–2015). Many, but not all measures, were based on HEDIS definitions. We did not include 2013 data due to the transition in the registry vendor, during which Advocate did not receive complete registry data. In 2014 and 2015, the data included detailed patient-level data with over 270 million unique patient measures. All physician-level data was aggregated up from individual patient-level data in 2014 and 2015. In addition to patient-physician quality information, we linked each physician to a 2015 physician characteristics database provided by Advocate that included demographic information such as age, gender, and board certification; group-level data including registry login rates; and patient panel data including number of chronic condition patients. Registry login rates were aggregated to the practice level because user ids were shared amongst clinicians and other staff members.

Our final dataset included physicians affiliated with Advocate from 2010 through 2015 and ten HEDIS-based measures that existed in the registry data across the full study period.[11] We constructed consistent measures across all years using patient-level quality data in 2014 and 2015, excluding measures when not collected in all years.[12] We also excluded the four ischemic vascular disease measures (blood pressure measurement, blood pressure control, anti-platelet medication, and body mass index screening) that had extremely high average physician baseline performance (in the range of 92% and 97%) due to ceiling effects.[13]

Study Design

We used an observational, quasi-experimental methodology that compared the observed performance of physicians to estimated performance in the absence of the registry, computed using a predictive piecewise model at the physician-measure level. This approach was similar to that of Campbell and colleagues (2007) who estimated the effect of the introduction of pay-for-performance in the UK's National Health Service (NHS) for primary care physicians. Importantly, like Campbell et al., we included physician fixed effects so that predicted performance improvement represents the average improvement within a physician rather than across physicians. The inclusion of physician fixed effects accounts for time-invariant physician characteristic confounders. Using a generalized linear model with a logit link function and normal distribution, we regressed yearly physician performance on a continuous year variable, an indicator for post-intervention period, and their interaction, as well as a fixed effect for each physician. The coefficients on the post-intervention period and the interaction variables indicated whether the introduction of the registry affected performance. We constructed models for each performance measure separately. The physician performance measure was computed as the proportion of patients within a physician's panel who met the national measure-specific benchmarks. Following Campbell and colleagues (2007), we adjusted our outcome variable using an empirical logit ($\log((p + .5 \times 1/n)/(1 - p + .5 \times 1/n))$ (where p is performance and n is measure panel size) to account for the likely non-linearity of this outcome as well as measurement error for small panel sizes.[14] We weighted each observation by a physician's average panel size for that measure over 6 years, doing so to minimize measurement error due to small panel sizes. Finally, we adjusted all p values for multiple comparisons using the Holm-Bonferroni method.

Additionally, we separately examined physicians who were low performing prior to the registry introduction as these physicians had the greatest ability to improve. To test whether low performing physicians had a different pattern of quality performance change, we applied the same piecewise regression methodology including only low performers. Low performers were identified as those providers with an average performance score in the bottom tertile across all measures and years 2010–2012 within their specialty.[15] Finally, we compared detailed physician characteristics for high and low improving physicians. We defined improvement as the difference between a physician's predicted and actual measure performance after the introduction of the registry, which can be thought of as improvement associated with the registry.[16] Top and bottom improvers were defined as physicians in the top and bottom tertile of improvement within their specialty.[17]

As a sensitivity analysis, we repeated the main analysis on the 49 physicians who participated in APP QI "collaboratives" between 2010 and 2012.

RESULTS

Study Sample

The study population included 434 physicians who were affiliated with Advocate between 2010 and 2015 and had at least one quality measure during this time. Physicians were 59% male (63% male in Illinois),[18] with 95% board certified and mean practice size of 3 physicians (Table 1). Just over 50% of the physicians were primary care physicians, about one-third were pediatricians, and the rest were cardiologists or endocrinologists. Physician panels included higher Medicare insurance coverage and lower Commercial and Medicaid relative to the state average.[19] Physician practices accessed the registry on average 9.2 times per week in 2015.

Registry Effects—Full Sample

The introduction of the registry was associated with performance improvement for two measures and performance decline for three measures (Table 2). A fifth measure was marginally associated with performance decline in the first year. The registry was not associated with a performance change for the five final measures. The two measures with improvement included childhood immunization status—rotavirus (9 percentage point and 8 percentage point improvement in years 1 and 2 ($p < 0.001$ for both)) and diabetes care—medical attention for nephropathy (3 percentage point and 2 percentage point improvement in years 1 ($p = 0.011$) and 2 ($p = 0.024$)). The three measures with decline included childhood immunization status—influenza (6 and 12 percentage point decline in years 1 ($p = 0.030$) and 2 ($p < 0.001$)), diabetes care—HbA1c testing (2 and 5 percentage point decline in years 1 ($p = 0.012$) and 2 ($p < 0.001$)), and diabetes care—poor HbA1c control (10 and 15 percentage point decline in years 1 and 2 ($p < 0.001$ for both)). The measure with marginally significant decline was diabetes care—eye exams (7 percentage point decline in year 1 ($p = 0.052$)). The four remaining measures where performance change was not associated with the registry introduction were childhood immunization status—combo 3, diabetes care—HbA1c control and both congestive heart failure (CHF)-appropriate medication measures—beta blockers and ACEi or ARB.

Table 1 Summary Statistics of Physicians in Study

Measure	Mean	Std dev	Min	Max
Average composite score	0.763	0.163	0	1
Age	55.228	9.119	36	78
Percent male	0.585	0.493	0	1
Years with Advocate	12.726	5.566	5	38
Years since residency	22.586	10.825	5	110
Board certified	0.947	0.224	0	1
Group size	3.134	2.526	1	10
Specialty				
Pediatrics	0.332			
Internal medicine	0.295			
Family medicine	0.216			
Cardiology	0.135			
Endocrinology	0.023			
Percent commercial	0.502	0.289	0	1
Percent Medicare	0.337	0.332	0	1
Percent Medicaid	0.129	0.227	0	1
Children—average panel size	7.452	18.571	0	238
Ischemic vascular disease—average panel size	11.528	16.442	0	106
Diabetes—average panel size	19.454	28.146	0	216
CHF—average panel size	0.882	1.595	0	14
Average number of Cerner Clicks/week (2015)*	9.158	6.853	1.058	34.173

The Composite score comes from physician-measure data between 2010 and 2015. All other data are from Advocate's 2015 internal data on physician characteristics

**Data on Cerner Clicks/week was aggregated to the physician group rather than the individual physician level because support staff share a login id with the physician and can login using multiple physician ids*

Registry Effects—Low Performing Physicians

Examining low performing physicians, the introduction of the registry was associated with performance improvement for one measure and performance decline for three measures (Table 3). One additional measure was associated with improvement in the first year and decline in the second year. The measures with associated improvement and decline were identical to the results with all physicians except for diabetic care—medical attention for nephropathy, which was no longer significant. The magnitude of the performance changes was smaller for low performing physicians relative to all physicians. For example, performance on childhood immunization status—rotavirus in year 2 had an 8 percentage point improvement for all physicians ($p \leq 0.001$) and 0.6 percentage point improvement for low performing physicians ($p = 0.021$). Performance on diabetes care—HbA1c testing in year 2 had a 5 percentage point decline for all physicians ($p \leq 0.001$) and a 2 percentage point decline for low performing physicians ($p = 0.005$).

Characteristics by Physician Improvement

The composite performance score prior to the registry introduction was significantly lower for physicians who improved the least relative to those who improve the most, 73% and 77% respectively (p value = 0.031) (Table 4). Physicians with larger improvements had a number of small differences in physician characteristics—they were younger (54.4 vs. 55.9, p value = 0.042), more likely to be in a larger group (3.5 vs 2.7 physicians, p value = 0.081), had a shorter tenure with APP (12.3 vs. 13.2 years, p value = 0.017), and fewer years since residency (21.4 vs. 23.6 years, p value = 0.072). Physician age, board certification, and insurance panel composition were not statistically different across groups. The panel composition by patient condition for top and bottom tertile physicians was different along all measures examined. Physicians who improved had more pediatric patients (7.8 vs. 6.2, p value < 0.001) and fewer ischemic vascular disease (10.0 vs. 11.9, p value < 0.001), diabetic (15.6 vs. 22.4, p value < 0.001), and congestive heart failure patients (0.8 vs. 0.9, p value < 0.001). Finally, groups with physicians who improved the most accessed the registry more frequently than those who improved the least—8.0 vs. 10.0 times per week ($p = 0.010$).

DISCUSSION

In this study of a natural experiment that enabled real-time performance feedback in an existing P4P program, registry

Table 2 Estimated Performance With and Without Registry for All Physicians

	Trend	Year 1 (2014) performance				Year 2 (2015) performance			
	2010–2012	With registry	Without registry	Year 1 effect	*P* value	With registry	Without registry	Year 2 effect	*P* value
Childhood imm. status									
Combo 3	− 0.010	0.804	0.815	− 0.010	0.598	0.851	0.804	0.047	0.188
Influenza	0.051	0.695	0.756	− 0.061	0.030	0.682	0.797	− 0.115	< 0.001
Rotavirus	0.021	0.969	0.881	0.088	< 0.001	0.974	0.898	0.076	< 0.001
Diabetes care									
HbA1c testing	0.013	0.931	0.954	− 0.023	0.012	0.917	0.963	− 0.045	< 0.001
Poor HbA1c control*	0.033	0.804	0.899	− 0.095	< 0.001	0.773	0.921	− 0.147	< 0.001
HbA1c control (< 8%)	0.000	0.709	0.724	− 0.015	1.000	0.676	0.725	− 0.048	0.369
Eye exams	0.011	0.594	0.660	− 0.066	0.052	0.601	0.670	− 0.069	0.155
Nephropathy†	0.018	0.939	0.911	0.029	0.011	0.948	0.924	0.024	0.024
CHF‡									
Beta blockers	0.034	0.806	0.826	− 0.020	1.000	0.823	0.852	− 0.029	0.628
ACEi or ARBs	0.033	0.792	0.813	− 0.021	1.000	0.852	0.840	0.012	0.681

All values come from the predictive piecewise model using an empirical logit transformation (see Online Appendix Section A3 for details). The P value tests whether the difference between with and without registry performance values is statistically significantly different from zero. Additionally, the P value has been adjusted for multiple comparisons using the Holm-Bonferroni method

**Measure is inverted for interpretability. A decrease in the proportion of diabetic patients with HbA1c > 9 or untested represents improvement*

†Medical attention for nephropathy

‡Congestive heart failure

implementation was not associated with systematic patterns of improved or worsened performance. Physicians with greater opportunity to improve, i.e., lower performing physicians at baseline, did not improve more than higher performing physicians, which is unlike findings in other P4P settings.[7]

Analysis of physician characteristics demonstrated that a number of characteristics were associated with greater improvement including larger group size, younger age, more pediatric patients, and fewer diabetic and congestive heart failure patients. Additionally, high improving physicians were in groups that accessed the registry more frequently than low improving physicians suggesting that registry access may be associated with performance improvement.

Our study has several limitations. First, we cannot definitely conclude that the performance changes in 2014 and 2015 we estimate were due solely to the introduction of the real-time registry. Practices and health systems have continually changing QI activities. However, APP did not run any QI programs outside of their registry education coincidental with the implementation of the new registry. Additionally, we separately

Table 3 Estimated Performance With and Without Registry for Low Performing Physicians

	Trend	Year 1 (2014)				Year 2 (2015)			
	2010–2012	With registry	Without registry	Year 1 effect	*P* value	With registry	Without registry	Year 2 effect	*P* value
Childhood imm. status									
Combo 3	− 0.026	0.665	0.694	− 0.029	1.000	0.789	0.739	0.059	0.102
Influenza	0.079	0.707	0.638	0.069	0.192	0.526	0.542	− 0.016	0.005
Rotavirus*	0.041	0.762	0.726	0.036	0.027	0.972	0.966	0.006	0.021
Diabetes care									
HbA1c testing	0.027	0.949	0.933	0.016	0.286	0.878	0.899	− 0.020	0.005
Poor HbA1c control†	0.044	0.882	0.852	0.031	< 0.001	0.719	0.747	− 0.029	< 0.001
HbA1c control (< 8%)	0.012	0.681	0.669	0.011	0.879	0.616	0.663	− 0.047	1.000
Eye exams	0.049	0.719	0.675	0.044	0.176	0.556	0.567	− 0.011	0.038
Nephropathy‡	0.037	0.914	0.889	0.024	0.245	0.927	0.926	0.001	1.000
CHF§									
Beta blockers	0.061	0.790	0.742	0.048	1.000	0.851	0.789	0.062	1.000
ACEi or ARBs	0.076	0.827	0.773	0.055	1.000	0.857	0.801	0.056	0.584

This table only includes physicians whose composite performance during years 2010–2012 were in the first (lowest) tertile of performance during that time within their specialty. All values are predicted probabilities from the piecewise model using an empirical logit transformation (see Online Appendix Section A3 for details). The P value tests whether the difference between with and without registry performance values is statistically significantly different from zero. Additionally, the P value has been adjusted for multiple comparisons using the Holm-Bonferroni method

**No observations existed in the bottom tertile for childhood rotavirus vaccination in 2014*

†Measure is inverted for interpretability. A decrease in the proportion of diabetic patients with HbA1c > 9 or untested represents improvement

‡Medical attention for nephropathy

§Congestive heart failure

Table 4 Characteristics of Bottom and Top Performers (Bottom and Top Tertile)

	Bottom performing physicians	Top performing physicians	*P* value
Average composite score prior to registry	0.730	0.774	0.031
Average improvement associated with registry*	− 0.001	0.110	< 0.001
Age	55.918	54.435	0.042
Percent male	0.623	0.529	0.243
Years with Advocate	13.212	12.333	0.017
Years since residency	23.582	21.351	0.072
Board certified	0.938	0.957	0.268
Group size	2.692	3.529	0.081
Percent commercial	0.471	0.507	0.357
Percent Medicare	0.362	0.309	0.151
Percent Medicaid	0.145	0.151	0.622
Children—average panel size	6.233	7.841	< 0.001
Ischemic vascular disease—average panel size	11.904	10.036	< 0.001
Diabetes—average panel size	22.390	15.645	< 0.001
CHF—average panel size	0.877	0.804	< 0.001
Number of Cerner Clicks/week	8.031	9.990	0.010

Composite score comes from physician-measure data between 2010 and 2012. The average improvement comes from the predicted piecewise model difference between the performance with and without the registry. All other data are from Advocate's 2015 internal data on physician characteristics. P values are from a two-sample t test with equal variance except for variables percent male and board certification, which come from a two-sample test of proportions

**Defined as the average across all physician's "improvement" or mean difference in a physician's actual performance and predicted performance without the registry*

analyzed the 49 physicians who participated in the APP "collaboratives" between 2010 and 2012 and found these physicians to have similar results (see Online Appendix Table A6).[20] APP was unaware of specific QI activities in individual practices; nonetheless, we cannot rule out the possibility of some activities affecting our results. Second, we compared observed scores with predicted scores (projections as if the registry had not existed) for physicians in the sample based on historic trends. Using a hypothetical comparison group may not accurately reflect the scores without the registry. Although an appropriate comparison group was not available due to Advocate's instantaneous rollout across all its physicians, we used a validated approach using a piecewise model based on the study of a P4P intervention in the UK where similar limitations on lack of available comparison groups existed.[2] Third, the data did not identify the individual physician or staff member accessing the registry, preventing additional analysis on the correlation between registry access and performance. Fourth, our results may not be generalizable to programs that are average or low performing. Fifth, the high baseline performance of Advocate physicians may have limited opportunity to improve because of a "ceiling effect". We attempted to address this by excluding measures with near 100% baseline performance and using an empirical logit transformation. Additionally, physicians in this study have been with APP for over a decade on average and may be less likely to modify their behavior relative to newer members. Similarly, physicians are in relatively small practices and may experience little peer pressure relative to larger groups, which could dampen intrinsic incentives. Finally, we were unable to use the full set of quality metrics Advocate collects due to changes in the measures collected over time.

A set of literature outside of P4P examines physician response to the public reporting of information without any link to financial incentives and has generally found physicians increase quality in response to the introduction of this information.[21–23] Most literature examining effects of P4P evaluates the introduction of a new program. This study contributes to a small body of literature that examines the role of a specific P4P mechanism: information. Out of the three P4P studies that identify the role of information on quality measure performance, two find no effect and one finds a partial effect.[24–26] However, the information provided in the studies was limited—two studies presented feedback twice a year and the third study presented clinic level feedback for a single measure once a week. This study is the first to examine the effect of providing access to real-time physician and patient-level information. Furthermore, we do not know of work that directly looks at changes in information levels (quarterly to real-time) rather than at the introduction of information.

Overall, the study suggests that providing additional information on performance in an environment with a fixed incentive structure is not sufficient to consistently change performance. On the other hand, the considerable complexity and disruptions to workflow did not lead to a systematic decline in quality. Furthermore, physician practices used the registry frequently, accessing the registry more than once a day.

The lack of uniform findings suggests that interventions that solely target "intrinsic" incentives may not be sufficient to increase performance, even if the intervention is well used. It is important to keep in mind the context of this study, a high performing, health system–affiliated set of physician groups. It may be harder for high performing groups to continually improve and affiliated rather than employed physicians may have weaker incentives to change behavior. This finding may be of particular relevance to high performing physician groups that already employ a strong "extrinsic" incentive—on average, physicians received a $10,000 P4P bonus per year and could receive up to $50,000 per year just for Advocate patients on their panels. Similar "intrinsic" interventions in average or low performing groups or groups with a smaller "extrinsic" incentive structure may produce different physician performance responses. Future research may consider other levers to enhance "intrinsic" motivation, such as EHR nudges or the use of comparative data (e.g., in the form of rankings) rather than individual data, or the direct pairing of new information with "extrinsic" or financial incentives.

Corresponding Author: *Amelia M. Bond, PhD; Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA (e-mail: amb2036@med.cornell.edu).*

Funding Information *This research was funded by the Robert Wood Johnson Foundation and Commonwealth Foundation.*

Compliance with Ethical Standards:

Conflict of Interest: *Dr. Bond receives research funding from Blue Cross Blue Shield of Louisiana, which does not have relationship to this manuscript. Dr. Emanuel has received speaking fees from various companies, organizations, and professional health care meetings. He has stock ownership in Nuna and is an investment partner in Oak HC/FT, neither of which have relationship to this manuscript. Dr. Navathe receives research funding from Hawaii Medical Services Association and Oscar Health Insurance. He also serves as an advisor to Navvis and Company, Navigant Inc., Lynx Medical, Indegene Inc., and Sutherland Global Services and receives an honorarium from Elsevier Press, none of which have relationship to this manuscript. All remaining authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Rosenthal MB**, **Frank RG**, **Li Z**, **Epstein AM**. Early experience with pay-for-performance: from concept to practice. JAMA. 2005;294:1788–93.
2. **Campbell S**, **Reeves D**, **Evangelos E**, **Middleton E**, **Sibbald B**, and **Roland M**. Quality of primary care in England with the introduction of pay for performance. New Engl J Med. 2007;357: 181–190.
3. **Rosenthal MB**. Beyond pay for performance–emerging models of provider-payment reform. New Engl J Med. 2008;359:1197–200.
4. **Ryan AM**, **Damberg CL**. What can the past of pay-for-performance tell us about the future of Value-Based Purchasing in Medicare? Healthc (Amst) 2013;1: 42–49.
5. **Findlay S.** Implementing MACRA. Health Affairs Health Policy Brief. Available at: https://www.healthaffairs.org/do/10.1377/hpb20170327.272560/full/. Accessed February 6, 2019
6. **Eijkenaar F**, **Emmert M**, **Scheppach M**, **Schöffski O**. Effects of pay for performance in health care: A systematic review of systematic reviews. Health Policy 2013;110(2–3): 115–130.
7. **Greene J**, **Hibbard JH**, **Overton V**. Large performance incentives had the greatest impact on providers whose quality metrics were lowest at baseline. Health Aff 2015;34(4):673–680.
8. Physicians directly receive their bonus checks during an annual APP meeting. For additional information on Advocate's CI Program see **Marcotte L, Hodlofski A, Bond A, Patel P, Sacks L, Navathe AS**. Into practice, Advocate Health System uses behavioral economics to motivate physicians in its incentive program. Healthc (Amst) 2017;5(3)129–135. https://doi.org/10.1016/j.hjdsi.2016.04.011
9. APP affiliated physicians choose their own EMRs. Therefore, the registry system is a way to link affiliated physicians in an organization that does not have consistent EMR systems.
10. Personal communication with a quality improvement leader in Advocate Physician Partners.
11. The two Chronic Heart Failure (CHF) measures are HEDIS based measures applied to a different population.
12. Two measures, Childhood Rotavirus Vaccination and Diabetes HbA1c < 8, were collected only between years 2011 and 2015.
13. Specifically, we excluded measures that were estimated to be above 100% in 2014 in our linear model. See the Study Design section for our more general regression framework and Online Appendix Table A4 for the linear model results.
14. Using a traditional logit transformation accounts for the likelihood that improving from 50% to 60% was easier than improving from 90% to 100%. Using the empirical logit rather than a pure logit both preserves all observations (a value of 0 and 1 are undefined as logit values) and shifts observations, particularly small panel sizes, towards 0.5.
15. Specifically, we calculated an average composite performance score at the physician-year level weighting each measure equally and then computed the composite score average across the years. Categorization of physician performance was based on composite performance within a specialty because many specialties only had certain measures. Categorizing across rather than within specialties would place the majority of one specialty in a single high/low performance category. For example, pediatricians typically had childhood immunization status measures, but no diabetic and CHF measures. Alternatively, cardiologists only had CHF measures.
16. Specifically, we used the GLM piecewise models to predict performance without the registry and calculated the mean difference in predicted and actual performance across a physician's measures.
17. Categorization of performance improvement was based on composite performance within a specialty because many specialties only had certain measures. See Footnote 13 for details.
18. Kaiser Family Foundation. Distribution of physicians by gender. Available at: http://www.kff.org/other/state-indicator/physicians-by-gender/. Accessed 6 February 2019.
19. Kaiser Family Foundation. Health insurance coverage of the total population. Available at: http://www.kff.org/other/state-indicator/total-population. Accessed 6 February 2019.
20. The magnitudes of the effect sizes are similar, however the results are not as significant. It is hard to detect meaningful differences with a much smaller sample size.
21. **Kolstad JT**. Information and quality when motivation is intrinsic: Evidence from Surgeon Report Cards. Am Econ Rev 2013;103(7):2875–2910.
22. **Schneider EC**, **Epstein AM**. Influence of cardiac-surgery performance reports on referral practices and access to care – A survey of cardiovascular specialists. New Engl J Med 1996;335(4): 251–256.
23. **Schneider EC**, **Epstein AM**. Use of public performance reports: A survey of patients undergoing cardiac surgery. JAMA. 1998;279: 1638–1642.
24. **Fairbrother G**, **Hanson K**, **Friedman S**, **Butts G**. The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates. Am J Public Health 1999;89(2): 171–175.
25. **Hillman AL**, **Ripley K**, **Goldfarb N**, **Weiner J**, **Nuamah I**, **Lusk E**. The Use of Physician Financial Incentives and Feedback to Improve Pediatric Preventive Care in Medicaid Managed Care. Pediatrics. 1999;104(4): 931–935.
26. **Roski J**, **Jeddeloh R**, **An L**, **Lando H**, **Hannan P**, **Hall C**, **Zhu SH**. The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines. Prev Med 2003;36(3): 291–299.