Vol. XLII, No. 2
March 2014

Changes in Health Care Financing & Organization (HCFO)

findings brief

key findings

- Study results indicate that neither the quality scorecard nor the quality incentive payment program had a significant positive effect on general clinical quality.
- Three main factors likely combined to weaken program effects: (1) modest size of the incentive; (2) use of rewards only; (3) targeting incentive payments to the group rather than to individuals.
- The researchers found that, relative to the scorecard and reporting alone, the addition of the Quality Incentive Payment Structure (QIP) was associated with a reduction in quality, a result contrary to the intent of the payment incentive program.

The Challenges in Achieving Successful P4P Programs

Overview

Health care payment reform is becoming one of the most important issues debated by health care policymakers, payers, providers, and purchasers. Architects of new payment models point out that the traditional fee-for-service model encourages unnecessary medications and procedures while capitation promotes limits on care and poses financial challenges to smaller provider groups.

During the late 1990s and early 2000s, pay-for-performance (P4P) programs grew in popularity. By design, P4P incorporates rewards for providing guideline-based services that mitigate the tendency toward underuse inherent in capitation and discourage fee-for-service–type overuse of expensive services—for example, through incentives for generic drug prescribing and appropriate use of antibiotics and asthma controller medications.

In a HCFO-funded study, Douglas A. Conrad, Ph.D., of the University of Washington School of Public Health and Community Medicine and colleagues examined a unique quasi-experiment that measured the effects of a large-scale P4P program implemented by a leading health insurer in Washington State between 2003 and 2007.[1] In its phased experiment, the plan recruited medical group practices and restricted the program to the products offered by commercial preferred provider organization plans. The researchers examined the clinical quality performance of three sets of medical groups: (1) those participating only in a quality scorecard (QSC) and public reporting program, (2) those participating in a quality incentives program (QIP) comprised of P4P payments in addition to the quality scorecard and reporting, and (3) a "control" group of roughly comparable practice organizations not participating in either the QSC or QIP program.

Experiment

Using a phased approach to conduct the experiment, the health plan first designed a quality scorecard that it pretested with an initial cohort between July 2001 and June 2002. Three medical groups started using the scorecard in 2002, and four additional groups began using it between 2003 and 2007. The initial three groups became eligible

Robert Wood Johnson Foundation

Changes in Health Care Financing and Organization is a national program of the Robert Wood Johnson Foundation administered by AcademyHealth.

for incentive payments starting in 2004, and the remaining four groups began participating in the QIP between 2004 and 2007. The health plan did not randomly select medical groups for the experiment but instead targeted a specific set of large medical groups for participation.

The medical leaders of the health plan selected a set of well-established metrics for the quality scorecard and provider incentive payment program:

- Breast cancer screening (mammogram) for women age 52-69 in the year prior to or during the measurement year.
- Cervical cancer screening (Pap test) for women age 21-64 in the 2 years prior to or during the measurement year.
- Well-child visits: 6 or more by age 15 months.
- Use of optimal medications for asthma: ages 5-56.
- Diabetes: 2 Hemoglobin A1c (HbA1c) tests during the measurement year.
- Diabetes: ACE-Inhibitor or ARB medication prescribed during the measurement year.
- Coronary artery disease: LDL screening during the measurement year.

The health plan structured the quality payment incentive on points for each measure. It incorporated both the level of achievement and degree of improvement from the previous year. During the first two years of the experiment (2003 and 2004), only the highest-scoring groups received incentive payments. This "contest" resulted in payments based on the relative performance of the participating groups. During the second two years of the experiment (2004 and 2005), incentive payments were based on how closely the medical groups came to reaching the achievable benchmarks of care and the extent of performance improvement over the previous year. The groups bore no risk. Rather, incentive payments were "new money."

Analysis

Given the phased implementation of the QSC and QIP components, the researchers used a modified difference-in-differences methodology. They compared the seven intervention groups with five comparison groups (the control group), which were selected in collaboration with the health plan. Most of the groups in each cohort were physician-owned. Given the plan's structured approach to soliciting medical groups, the researchers used methods that mitigated potential selection bias but noted that they could not completely rule out factors that might confound their estimates.

The researchers developed and used a regression model to estimate the effects on quality of two health plan quality programs: the QIP and QSC. They included patient-level variables to control for factors that could affect providers' quality achievement scores.

Results

To detect differential patterns in quality performance over time, the researchers constructed a time-series plot for each quality measure. They distinguished among three cohorts: (1) Ever QSC only, (2) Ever QSC/QIP, (3) and the control medical groups.

In general, the researchers did not find marked differential changes over time among cohorts on the quality indicators, although they observed some baseline differences including breast cancer screening levels.

Both the scorecard and incentive program cohorts showed considerably greater achievement over time in LDL cholesterol screening among diabetes patients within the intervention groups than within the controls. The intervention cohort also showed some quality performance improvement in ACE-inhibitor use among diabetics. The researchers noted that because both intervention cohorts started lower at baseline, they had more room for improvement. Overall, the descriptive time series plots analyzed by the researchers failed to reveal any added benefit of program participation beyond sentinel effects.

In their analysis of the effect of the payment incentives on the quality measures, the researchers found that neither the scorecard and reporting alone nor the QIP incentive had a positive effect on quality. The analysis of the plan's experiment revealed results that were opposite to the intent of the payment incentive program.

Discussion and Policy Implications

In considering their null findings, the researchers pointed to several observations, some of which were drawn from key informant interviews. They noted that the modest size of the incentive payment likely played a role in the results. However, the study did not indicate a "treating to the test" phenomenon associated with targeting certain measures; the non-incentivized services did not appear to be negatively affected. The concentration on patients from one health plan could have limited the success of the experiment, although studies in California have shown similar results with multi-payer experiments. Some physician interviewees noted that the production-based incentives were not aligned with the goal of the payments to improve quality. Physicians also felt that a significant drawback of the experiment was the group rather than individual nature of the incentive program.

The researchers made additional observations. The health plan structured the experiment to transition from a relative performance model to an absolute performance standard. There was no evidence to show this shift generated positive results, possibly due to the lack of any downside risk. Yet, while penalties may have been a stronger motivation to perform, this structure would need to be weighed against the negative reactions to reducing a physician's income.

This study offers five major contributions to the larger body of literature on P4P: (1) The results of this study support other research showing that P4P has a very small impact on quality. (2) This is the first P4P study to contrast the effects of quality incentives based on relative performance versus absolute performance based on achievable benchmark standards within the same study sample. (3) The phased-in nature of the publicly reported quality scorecard, followed by implementation of the quality incentive program, allows one to separate the effects of the scorecard from those of the scorecard plus explicit quality incentives. (4) The study combines key informant interview data with the quantitative results to provide a richer interpretation of the findings. (5) This research is also one of the few P4P studies to explicitly control for case mix differences.

Conclusion

Given the modest success of many P4P studies, it seems that other means of controlling costs and increasing quality should be explored. The researchers call for "a full-court press on quality and efficiency, based on common and broadly defined clinical and economic metrics among multiple payers and providers."

For more information

Contact Douglas A. Conrad, Ph.D., at dconrad@u.washington.edu.

About the author

Emily Blecker, B.A., is a research assistant at AcademyHealth with the Changes in Health Care Financing and Organization (HCFO) initiative. She may be reached at 202-292-6736 or at emily.blecker@academyhealth.org.

Endnotes

1. For complete findings, see Conrad, D.A., Grembowski, D., Perry, L., Maynard, C., Rodriguez, H., and Martin, D., Paying physician group practices for quality: A statewide quasi-experiment, *Healthcare*, Vol. 1, No. 3-4, 2013, pp 108-116.