QUERI

Understanding the Intervention and Implementation Factors Associated with Benefits and Harms of Pay for Performance Programs in Healthcare

May 2015

Prepared for:
Department of Veterans Affairs
Veterans Health Administration
Quality Enhancement Research Initiative
Health Services Research & Development Service
Washington, DC 20420

Prepared by:
Portland Evidence-based Synthesis Program (ESP) Center
Portland VA Medical Center
Portland, OR
Devan Kansagara, MD, MCR, Director

Investigators:
Principal Investigator:
Karli Kondo, PhD

Contributing Investigators:
Cheryl Damberg, PhD, MPH
Aaron Mendelson, BA
Makalapua Motu'apuaka, BS
Michele Freeman, MPH
Maya O'Neil, PhD
Rose Relevo, MLIS, MS
Devan Kansagara, MD, MCR

VA HEALTH CARE Defining EXCELLENCE in the 21st Century

PREFACE

Quality Enhancement Research Initiative's (QUERI) Evidence-based Synthesis Program (ESP) was established to provide timely and accurate syntheses of targeted healthcare topics of particular importance to Veterans Affairs (VA) clinicians, managers and policymakers as they work to improve the health and healthcare of Veterans. The ESP disseminates these reports throughout the VA, and some evidence syntheses inform the clinical guidelines of large professional organizations.

QUERI provides funding for four ESP Centers and each Center has an active university affiliation. The ESP Centers generate evidence syntheses on important clinical practice topics, and these reports help:

- develop clinical policies informed by evidence;
- guide the implementation of effective services to improve patient outcomes and to support VA clinical practice guidelines and performance measures; and
- set the direction for future research to address gaps in clinical knowledge.

In 2009, the ESP Coordinating Center was created to expand the capacity of HSR&D Central Office and the four ESP sites by developing and maintaining program processes. In addition, the Center established a Steering Committee comprised of QUERI field-based investigators, VA Patient Care Services, Office of Quality and Performance, and Veterans Integrated Service Networks (VISN) Clinical Management Officers. The Steering Committee provides program oversight, guides strategic planning, coordinates dissemination activities, and develops collaborations with VA leadership to identify new ESP topics of importance to Veterans and the VA healthcare system.

Comments on this evidence report are welcome and can be sent to Nicole Floyd, ESP Coordinating Center Program Manager, at Nicole.Floyd@va.gov.

Recommended citation: Kondo K, Damberg C, Mendelson A, Motu'apuaka M, Freeman M, O'Neil M, Relevo R, Kansagara D. Understanding the intervention and implementation factors associated with benefits and harms of pay for performance programs in healthcare. VA-ESP Project #05-225; 2015.

This report is based on research conducted by the Evidence-based Synthesis Program (ESP) Center located at the VA Portland Health Care System, Portland, OR, funded by the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Quality Enhancement Research Initiative. The findings and conclusions in this document are those of the author(s) who are responsible for its contents; the findings and conclusions do not necessarily represent the views of the Department of Veterans Affairs or the United States government. Therefore, no statement in this article should be construed as an official position of the Department of Veterans Affairs. No investigators have any affiliations or financial involvement (*eg*, employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in the report.

ACKNOWLEDGMENTS

The authors would like to thank the Key Informants and Technical Expert Panel for their participation and valued input. In particular we would also like to recognize and thank Laura Damschroder, MS, MPH for her contribution the conceptual framework used in this review.

EXECUTIVE SUMMARY

INTRODUCTION

Over the last decade, pay for performance (P4P) programs have been implemented in a variety of health systems, including the VHA, as a means to improve the efficiency and quality of health care. There has been a parallel increase in the number of studies examining the effects of P4P. A number of recent reviews have summarized this literature, but have generally found insufficient evidence to broadly characterize the balance of harms and benefits. However, financial incentives programs are complex interventions whose effects may depend in part on the settings in which they are implemented, the methods used for implementation, the populations targeted, and the characteristics of the incentive programs themselves.

The objectives of this report are to summarize the positive and negative effects of P4P on process and health outcomes, and to examine how implementation characteristics modify the effects of P4P programs. The Key Questions used to guide our report are:

Key Question 1: What are the effects of pay for performance programs on patient outcomes and processes of care?

Key Question 2: What implementation factors modify the effectiveness of pay for performance?

Key Question 3: What are the positive and negative unintended consequences, including any effect on health disparities, associated with pay for performance?

METHODS

A comprehensive, good-quality systematic review on Value-Based Purchasing, including P4P programs, was released by the RAND Corporation in March 2014. We searched PubMed, PsycINFO (Ovid), and CINAHL (EBSCO), and limited our search to include studies published in the time period between the end of their search date and April 2014, and studies examining programs not included in the RAND report (*eg*, UK's Quality and Outcomes Framework [QOF]). We also conducted an internet (Google) search without date limits for unpublished literature using keywords included in our search strategy and targeting the names of specific P4P programs (*eg*, QOF, Hospital Quality Incentive Demonstration [HQID]), and we searched websites including the RAND Corporation, the Agency for Healthcare Research and Quality (AHRQ), and the National Institute for Health and Care Excellence (NICE).

We included studies evaluating P4P programs targeting healthcare providers at the individual, group, managerial, or institutional level. We included studies conducted in countries whose health systems are similar to portions of the US health system, and excluded pediatric populations. To assess the effects of P4P on process of care and health outcomes, we only included studies that enrolled more than 10,000 patients, included a comparison group, and/or conducted a time-series analysis. Studies with smaller patient samples and pre-post study designs were included to assess implementation characteristics and harms/unintended consequences. One investigator abstracted data and assessed study quality, with review by a second investigator. We qualitatively synthesized the results and organized them according to a model we adapted from existing P4P and implementation models.

In collaboration with the primary author, we provide a summary of RAND's findings on P4P programs relevant to the VHA. In addition, we engaged 14 experienced P4P researchers as key informants (KI) to gain insight into issues related to implementation and unintended consequences. We conducted hour-long semi-structured interviews with KIs to understand their perceptions of implementation factors that were important in influencing both the positive and negative impacts of P4P programs. Five investigators conducted independent inductive open-coding of interview notes. One investigator with qualitative research experience (KK) reviewed investigators' codes and identified common themes.

RESULTS

Of 1,363 titles and abstracts identified from the electronic search we reviewed the full text of 509 potentially relevant articles, and found 93 studies that met inclusion criteria. We included one additional article recommended by a peer reviewer, for a total of 94 included studies. We identified 47 primary studies for Key Question 1, 41 primary studies meeting inclusion criteria for Key Question 2, and 42 primary studies addressing Key Question 3. Thirty-two studies met criteria for more than one key question. These results include findings from our literature search and themes that emerged during our interviews with key informants. In addition to what is presented in this executive summary, the main report also includes a summary of RAND's key findings, written in collaboration with the report's primary author.

Key Question 1: What are the effects of pay for performance programs on patient outcomes and processes of care?

Overall, we found that P4P programs in ambulatory settings can improve the proportion of patients receiving the care process targeted by an intervention. However, we consider this low-strength evidence because of inconsistencies across studies, lack of impact over the long term, heterogeneity of interventions studied and outcomes measured, and the typically small effect size. Studies of the UK's Quality and Outcomes Framework (QOF) consistently report modest improvements in the first one to 2 years of the program, particularly in practices with initial lower levels of attainment, followed by either a plateau or slowing of improvement rates. A handful of studies, particularly those evaluating Taiwan's diabetes mellitus P4P program, report moderate short-term improvements in processes of care, screening rates, and provision of preventive care associated with P4P. However, findings from longer-term studies examining processes of care often report a slowing of improvement or little to no association.

There is no clear, consistent evidence of the QOF's effect on patient outcomes. Similar to the process of care outcome results, the QOF had an immediate positive effect on some patient outcomes, but the rate of improvement was not sustained over time. For others, such as HbA1c, post QOF trends were significantly below those predicted before the intervention. In other countries and in the United States, there is little good-quality evidence that directly examines the effects of P4P on health outcomes, with most studies reporting little to no effect.

In hospital settings, studies evaluating the Premier Hospital Quality Incentive Demonstration (HQID) and the Hospital Value-Based Purchasing (HVBP) programs in the United States report a limited effect on both processes of care and patient outcomes. However, a study evaluating the effect of P4P in the VHA on processes of care found significant and sustained improvement on 6 of the 7 measures examined. Internationally, studies evaluating hospital P4P programs report generally positive effects, with a slowing of improvements or a plateau over time.

Key Question 2: What implementation factors modify the effectiveness of pay for performance?

a. What implementation factors are associated with changes in processes of care or patient outcomes?

We found 28 studies examining factors associated with processes of care or patient outcomes. We provide a more detailed summary of study and relevant key informant interview findings organized according to subcategories of the implementation framework in Table 1 (definitions of the implementation framework components are provided in the main report).

b. What implementation factors are associated with changes in provider cognitive and/or behavioral responses?

We included 14 studies examining factors associated with changes in provider cognitive and/or behavioral outcomes. Studies reported that perceptions of program effectiveness were related to measure alignment with goals, and that providers placing a higher degree of importance on goals and quality targets performed better than those who did not. In addition, measures focused on patient care experience or clinical quality improved staff communication and care coordination, while those focused on productivity or efficiency were associated with poor staff communication. One study found that provider participation in P4P programs relates to both the potential for rewards as well as perceived ethical risk, and another found differences in performance by underlying payment structure and concluded that higher incentives may be necessary when the degree of cost sharing is lower. Finally, the results of 2 small studies that surveyed providers on attitudes and values found a negative relationship between performance and placing a high value on autonomy.

KI discussions in this area centered on the balance between intrinsic and extrinsic motivation for providers and the organizational culture and support to align the two, including provider buy-in, and supportive and encouraging communication and feedback on provider performance. In addition, KIs stressed the importance of implementation processes, for programs in general and also for the introduction of newly incentivized measures. Implementation processes should be transparent and provide resources to encourage and enable provider buy-in through information that allows them to link the measure to clinical quality and provides guidance on how to achieve success. To further achieve buy-in, KIs urged the engagement of stakeholders of all levels at each stage, and recommended a “bottom-up” approach to program development. They stressed that P4P programs should include a combination of measures addressing processes of care and patient outcome, and that while measures should cover a broad range, too many measures increase the likelihood of negative unintended consequences. KIs also agreed that measures should reflect organizational priorities, be realistically attainable, evidence-based, clear, simple, and linked to clinically significant rather than data-driven outcomes, with systems in place for evaluation and modification as needed. In addition, improvements should be incentivized, incentives should be large enough to provide motivation but not so large as to encourage gaming, penalties may be more effective than rewards, and team-based incentives were suggested to increase the buy-in and professionalism of both clinical and non-clinical staff. Similarly, the timing of payments should be frequent enough to reinforce the link between measure achievement and the reward. However, this must be balanced with payment size, as the reward must be substantial enough to reinforce behavior.

Table 1. KQ 2 Evidence and Policy Implications by Implementation Framework Category

Implementation Framework Category	Study Evidence	Themes from KI Interviews	Policy Implications
Program design features	Thirteen studies[2-7,8-14] examined program design features and found: · Measures linked to quality and patient care were positively related to improvements in quality and greater provider confidence in the ability to provide quality care, with measures tied to efficiency were negatively associated. · Perceptions of program effectiveness were related to the perception that measures aligned with organizational goals, and perceived financial salience related to measure adherence, as did perceptions of target achievability. · Different payment models result in differences in both bonuses/payments and performance · More statistically stringent methods of creating composite quality scores was more reliable than raw sum scores · The cost effectiveness of P4P varies widely by measure.	· Programs should include a combination of process of care and patient outcome measures. · Process of care measures should be evidence-based, clear and simple, linked to specific actions rather than complex processes, and clearly connected to a desired outcome. · Measure targets should be grounded in clinical significance rather than data improvement. · Disseminate the evidence behind and rationale for incentivized measures · Measures should reflect the priorities of the organization, its providers, and its patients. · Incentives should be designed to stimulate different actions depending on the level of the organization at which they are targeted. · Incentives must be large enough to motivate, and not so large as to encourage gaming - with hypotheses ranging from 5-15% · Incentives should be based on improvements, and all program participants should have the ability to earn incentives · Magnitude of the incentive attached to a specific measure should be relative to organizational priorities	· Programs that emphasize measures that target process of care or clinical outcomes that are transparently evidence-based and viewed as clinically important may inspire more positive change than programs that use measures targeted to efficiency or productivity, or do not explicitly engage providers from the outset. · The incentive structure needs to carefully consider several factors including incentive size, frequency, and target.
Implementation Processes	Eight studies[15-2021,22] examined changes in implementation, with 7 specifically related to updating or retiring measures, and found: · Under both the QOF and in the VHA,	· Stakeholder involvement and provider buy-in are critical · Bottom up approach · Reliable data/feedback to providers in a	· P4P programs should target areas of poor performance and consider de-emphasizing areas that have achieved high performance.

Implementation Framework Category	Study Evidence	Themes from KI Interviews	Policy Implications
	removing an incentive from a measure had little impact on performance once a high level performance had been achieved. · Increasing maximum thresholds resulted in greater increases by poorer performing practices.	non-judgmental fashion · Consider distributing incentives to clinical and non-clinical staff	
Outer Setting	Seven studies[10,23-28] examined implementation factors related to the outer setting. · There is no clear evidence that setting (*eg*, region, urban vs rural) or patient population predict P4P program success in the long term.	· Measures should be realistic within the patient population and health system in which they are used · Programs should be flexible to allow organizations to meet the needs of their patient populations	· P4P programs should have the capacity to change over time in response to ongoing measurement of data and provider input.
Inner Setting	Eighteen studies[7,24,26-41] examined implementation factors related to the inner setting. Studies found: · For providers, being a contractor rather than being employed by a practice was associated with greater efficiency and higher quality. · Under the QOF, practices improved regardless of list size, with larger practices performing better in the short term. · Under the QOF there is limited evidence that group practice and training status was associated with a higher quality of care. · Findings were less clear in the US and elsewhere with regard to practice size and training status.	· Resources must be devoted to implementation, particularly when new measures are introduced · Provide support at the local level including designating a local champion · Incentives are just one piece of an overall quality improvement program. Other important factors may include a strong infrastructure, organizational culture, allocation of resources, and public reporting · Public reporting is a strong motivator and future research should work to untangle public reporting from P4P	· Programs that emphasize measures that target process of care or clinical outcomes that are transparently evidence-based and viewed as clinically important may inspire more positive change than programs that use measures targeted to efficiency or productivity, or do not explicitly engage providers from the outset. · P4P programs should have the capacity to change over time in response to ongoing measurement of data and provider input.
Provider characteristics	Four studies[5,23,39,42] examined characteristics of the individuals involved, and provided no strong evidence that provider characteristics such as gender, experience, or specialty play a role in P4P program success.		

Note: Categories are not mutually exclusive

Key Question 3: What are the positive and negative unintended consequences associated with pay for performance?

Forty-two studies examining unintended consequences associated with P4P met inclusion criteria for Key Question 3, of which 33 evaluated the QOF. Among these studies, 28 of the 42 evaluated the effect of P4P on health disparities in populations of low socio-economic status or racial/ethnic minorities, or examined disparities associated with other characteristics such as age and multiple conditions. Nineteen studies report findings related to other unintended consequences, such as gaming, positive and negative effects on unincentivized areas of care, and cherry-picking/risk selection.

Health Disparities

Most of the studies examining differential effects of P4P by race/ethnicity, socioeconomic, or other demographic characteristics came from the UK's QOF program. In general, there was no strong consistent evidence that P4P had different effects on different patient subgroups, though there were some exceptions as detailed in the main report. Groups with lower baseline care quality tended to experience greater absolute levels of improvement over the short term.

Key informants in the UK noted that, in the first 2 years after its introduction, the QOF successfully decreased health disparities. This was due to the larger magnitude of improvements seen among practices in areas of high deprivation which tended to have lower baseline levels of performance. However, key informants also noted that once practices were performing near the upper thresholds, the costs associated with eliminating the remaining gaps were higher in areas with higher deprivation, and that providers in more affluent areas were more likely to receive incentives.

In the United States, the relationship between P4P and health disparities has not been well studied. A number of KIs stressed the lack of formal evaluation of health disparities in US programs, the importance of the collection of cultural variables to allow for an accurate assessment, and the need for consistency across measures to allow for formal evaluation.

Other Unintended Consequences

Gaming

We found very few studies which directly examined the issue of gaming. Two studies examined preferential recording of values within the QOF, with one study reporting an increase of values just below a newly introduced target, and another study reporting no evidence of gaming. Key informants stressed that gaming is likely to occur and that P4P programs should be designed with this assumption. In general, KIs felt that to reduce the likelihood of gaming P4P programs must have stakeholder input and buy-in, and should be based on precise, simple, evidence-based, and realistic measures.

Risk selection

A number of studies examined risk selection associated with the QOF. One study found a positive relationship between the rate of exception reporting and total QOF score, and another study found significantly higher levels of quality in patients who were not excluded as compared with all patients, particularly for more complex processes and treatment-related indicators.

Studies report higher rates of exception reporting for non-white, low-income patients, and patients with more co-morbid disorders, with one study reporting a higher percentage of excluded patients in larger practices. However, another concluded that higher rates of exception reporting were due to better documentation associated with the QOF. In Taiwan, non-enrolled patients were older, had more co-morbid conditions, and had higher diabetes risk scores. Key informants in the UK felt that exception reporting was not being abused. In the United States, key informants expressed concern that higher risk patients can now be easily identified using algorithms, and a common theme among KIs was that incentive payments should be risk-adjusted to account for higher-risk patients.

Spillover effects

We found evidence of both positive and negative impacts of P4P on unincentivized measures as well as on unincentivized populations. One QOF study found that, over 3 years, the rate of improvement in areas or populations not associated with incentives declined. However, other studies in both the UK and the US reported positive effects on unincentivized care. For example, one study reported a positive spillover of a 10.9% increase in the recording of unincentivized indicators for patients with targeted disease conditions in the QOF. Key informants agreed that spillover effects likely occur, and suggested that the lack of significant findings associated with Centers for Medicare and Medicaid Services' (CMS) Hospital Value-Based Purchasing (HVBP) program was due to improvements in quality spilling over to control hospitals.

DISCUSSION

We found 94 studies conducted in the United States and other countries that could inform practice in the VHA. The studies we examined across all 3 Key Questions differed widely by health system and patient population, and evaluated a range of P4P programs that varied substantially in both measures prioritized and incentive structure. Despite numerous examples of P4P programs, the heterogeneity inherent in each health system and organization and the challenges related to the evaluation of complex interventions such as P4P preclude us from drawing strong conclusions that can be broadly applied.

While the literature does not provide strong evidence to definitively guide the implementation of P4P programs, there are several themes from KI interviews that were consistent with evidence from the published literature. First, programs that emphasize measures that target process of care or clinical outcomes that are transparently evidence-based and viewed as clinically important may inspire more positive change than programs that use measures targeted to efficiency or productivity, or do not explicitly engage providers from the outset. Findings from both the literature examining physician perceptions and KI interviews support the use of evidence-based measures that are congruent with providers expectations for clinical quality, and there was a strong agreement among KIs that provider buy-in is crucial.

Second, the incentive structure needs to carefully consider several factors, including incentive size, frequency, and target. In general, the QOF, with its larger incentives, has been more successful than programs in the US Key informants attribute this to incentives that are large enough to motivate behavior, but also caution that larger incentives may not be cost effective and may result in gaming. KIs also stressed the importance of the attribution of the incentive to provider behavior, that incentivized measures should be congruent with institutional priorities,

address the needs of the institution at the local level, and should be designed to best serve the local patient population.

Third, P4P programs should have the capacity to change over time in response to ongoing measurement of data and provider input. Key informants strongly agreed that P4P programs should be flexible and evaluated on an ongoing and regular basis. They pointed to the QOF, which is evaluated annually, and which since its inception has undergone numerous adjustments, such as to the measures incentivized and the thresholds associated with payments.

Finally and relatedly, P4P programs should target areas of poor performance and consider de-emphasizing areas that have achieved high performance. Findings from studies of both the QOF and the VHA and our KI interviews support that improvements associated with measures achieving high performance can be sustained after the measure has been de-incentivized. Consistent evaluation of the performance of, and adjustments to, incentivized measures will allow institutions to shift focus and attention to the areas of greatest need for improvement.

Recommendations for Future Research

Despite numerous P4P programs in the United States, the United Kingdom, and elsewhere, there is a need for higher-quality evidence to better understand whether these programs are effective in improving the quality of healthcare and patient health, and whether they result in negative unintended consequences. Studies examining P4P have been largely observational and primarily retrospective, or lack good matched comparison groups. In addition, one of the fundamental challenges in evaluating complex multi-component interventions such as P4P is disentangling the individual effect of each intervention. In the case of P4P, the challenge is even greater, as contextual and implementation factors must also be strongly considered, as programs differ widely in their measures and incentive structures, as do the overarching health systems and organizations to which they are applied, and the patient populations for which they are designed to serve. There is an urgent need to examine the implementation factors that may mediate or moderate program effectiveness, such as the influence of public reporting, the number and focus of measures, incentive size, structure, and target. In addition, more research is needed to better delineate whether P4P differentially affects subpopulations of patients, and if so, how best to mitigate health disparities and to avoid unintended consequences. Finally, KIs stressed the belief that the VHA as a system is in a unique position from which to conduct much-needed rigorous and methodologically strong P4P research, examining not only P4P's effectiveness on processes of care and patient outcomes, but also examining implementation characteristics and unintended consequences.

Limitations

Our review has a number of limitations, which are detailed in the full report. These limitations relate to the heterogeneity of the literature itself, the quality of included studies, and the preponderance of data on ambulatory care programs rather than hospital-based programs.

Conclusions

In general, P4P programs appear to have the potential to improve process of care outcomes over the short term, especially in ambulatory settings. There is insufficient evidence that P4P programs have beneficial effects on care processes over the long term, or on patient outcomes

over any time period. Incentive programs tend to have the greatest absolute effect on care processes over the short term in settings with lower baseline levels of performance. In the United States in particular, the effects of P4P on health disparities are unclear, largely due to the lack of patient cultural variables collected and recorded. There is limited evidence in the QOF and VHA that initial improvements may be sustained even after removal of the incentive. The value of incentive programs to stimulate incremental performance gains once initial improvements have been achieved is unclear. Also unclear is the influence of P4P above and beyond other quality initiatives often accompanying financial incentives, such as public reporting and information technology. Findings from experts in the field are congruent with previous qualitative work – that the potential negative unintended consequences of P4P may outweigh benefits in these circumstances, though there is relatively little good-quality evidence examining the rates of harms from P4P.