REVIEW

Implementation Processes and Pay for Performance in Healthcare: A Systematic Review

Karli K. Kondo, PhD, MA[1,3], Cheryl L. Damberg, PhD, MPH[2], Aaron Mendelson, BA[3], Makalapua Motu'apuaka, BS[1], Michele Freeman, MPH[1], Maya O'Neil, PhD, MS[1,3], Rose Relevo, MLIS, MS[1], Allison Low, BA[1], and Devan Kansagara, MD, MCR[1,3]

[1]Portland Veterans Affairs Medical Center, Evidence-based Synthesis Program, Portland, OR, USA; [2]RAND Corporation, Santa Monica, CA, USA; [3]Oregon Health and Science University, Portland, OR, USA.

BACKGROUND: Over the last decade, various pay-for-performance (P4P) programs have been implemented to improve quality in health systems, including the VHA. P4P programs are complex, and their effects may vary by design, context, and other implementation processes. We conducted a systematic review and key informant (KI) interviews to better understand the implementation factors that modify the effectiveness of P4P.

METHODS: We searched PubMed, PsycINFO, and CINAHL through April 2014, and reviewed reference lists. We included trials and observational studies of P4P implementation. Two investigators abstracted data and assessed study quality. We interviewed P4P researchers to gain further insight.

RESULTS: Among 1363 titles and abstracts, we selected 509 for full-text review, and included 41 primary studies. Of these 41 studies, 33 examined P4P programs in ambulatory settings, 7 targeted hospitals, and 1 study applied to nursing homes. Related to implementation, 13 studies examined program design, 8 examined implementation processes, 6 the outer setting, 18 the inner setting, and 5 provider characteristics. Results suggest the importance of considering underlying payment models and using statistically stringent methods of composite measure development, and ensuring that high-quality care will be maintained after incentive removal. We found no conclusive evidence that provider or practice characteristics relate to P4P effectiveness. Interviews with 14 KIs supported limited evidence that effective P4P program measures should be aligned with organizational goals, that incentive structures should be carefully considered, and that factors such as a strong infrastructure and public reporting may have a large influence.

DISCUSSION: There is limited evidence from which to draw firm conclusions related to P4P implementation. Findings from studies and KI interviews suggest that P4P programs should undergo regular evaluation and should target areas of poor performance. Additionally, measures and incentives should align with organizational priorities, and programs should allow for changes over time in response to data and provider input.

Electronic supplementary material *The online version of this article (doi:10.1007/s11606-015-3567-0) contains supplementary material, which is available to authorized users.*

Published online March 7, 2016

KEY WORDS: pay for performance; financial incentives; implementation; performance metrics; systematic review.
J Gen Intern Med 31(Suppl 1):S61–9
DOI: 10.1007/s11606-015-3567-0

INTRODUCTION

Pay for performance (P4P) refers to the use of financial incentives to stimulate improvements in healthcare efficiency and quality. P4P belongs to a collection of financing schemes known as alternative payment models (APMs), which are designed to replace fee-for-service (FFS) payment systems. Whereas FFS payment rewards volume of services, APMs are designed to incentivize better outcomes and value. This is typically achieved by ensuring that providers and systems are financially vested in patient health status and efficient care delivery. In addition to P4P, prominent models include bundled payments and medical homes. Although P4P had previously been implemented by private payers on a small scale, there has been an increase in large-scale ambulatory and hospital P4P programs over the last decade both in the United States and internationally.

The Veterans Health Administration (VHA) instituted its performance pay program in 2004 after passage of the VA Health Care Personnel Enhancement Act.[1] The amount of performance pay awarded to each provider is determined by the degree to which they achieve a set of performance goals, which may include measures of care processes (e.g., ordering periodic hemoglobin A1c tests in diabetic patients), health outcomes, or fulfillment of work responsibilities (e.g., timely completion of training activities). There is also a managerial performance pay program for administrators. The VHA performance pay program allows medical centers and regional networks autonomy in determining the choice of measures comprising the performance goals for different types of providers. In 2011, approximately 80 % of VA providers received performance pay, at an average of $8,049 per provider.[2]

In recent years, there have been an increasing number of studies examining the effects of these and other large-scale P4P programs. As experience with and evidence in examining these programs have increased, questions have arisen regarding the effectiveness of such programs and concerns voiced about the potential for negative unintended consequences.[3,4]

However, financial incentive programs are complex interventions and vary widely in the their implementation, including characteristics of the measures chosen, such as the number of measures incentivized, the types of measure (e.g., structural, cost/efficiency, clinical processes, patient/intermediate outcomes, patient experience), and features related to the incentive structure such as who the incentive targets (e.g., providers, groups, managers, administration), the amount, whether incentives are in the form of rewards (e.g., fee differentials, bonuses) or penalties (e.g., withholding payment, repayments to payers), and incentive frequency. Added to the complexity are differences in the contexts in which they are implemented, such as the type of setting (e.g., ambulatory settings, hospitals, nursing home), the organizational culture within the setting, and other factors including patient population. The positive and negative effects associated with any given P4P program likely depend in part on the combination of all of these factors.

This paper, which is part of a larger report commissioned by the VHA, reports the results of a systematic review and key informant (KI) interviews focused on how implementation features influence the effectiveness of P4P programs.

METHODS

Data Sources and Strategy

A recent report on value-based purchasing published by the RAND Corporation included an examination of P4P programs.[5] We modified their search strategy and conducted an updated search of the PubMed, PsycINFO, and CINAHL databases from the end of their search date through April 2014. We searched the grey literature, targeting websites of both organizations known to conduct systematic reviews and those known to have experience or data related to P4P programs. In addition, we performed searches of PubMed and Google, targeting the names of larger P4P programs, and also searched for studies examining programs not included in the RAND report (e.g., the UK Quality and Outcomes Framework [QOF]) from database inception through April 2014 (Appendix 1, available online). We obtained additional articles from systematic reviews, reference lists of pertinent studies, reviews, and editorials, and by consulting experts.

Study Selection

We included English-language trials and observational studies examining direct pay-for-performance programs targeting healthcare providers at the individual, group, managerial, or institutional level. We excluded studies examining patient-targeted financial incentives, as well as payment models other than direct pay-for-performance, such as managed care, capitation, bundled payments, and accountable care organizations. Only studies examining systems and patient populations similar to that of the VHA were included, thus excluding studies conducted in countries with healthcare systems that differ widely from U.S. or VHA settings, studies that were not conducted in hospital or ambulatory settings, and studies with child patient populations (Appendix 2, available online). Two investigators independently assessed each study for inclusion based on the criteria (Appendix 3, available online). We used a "best evidence" approach to guide study design criteria, according to the question under consideration and the literature available.[6]

Data Extraction and Quality Assessment

We abstracted data from each included study on study design, sample size, country, relevance to the VHA, program description, incentive structure, incentive target (e.g., provider, management, administration), comparator, outcome measures, and results. Given the wide variation in study designs and large number of observational studies, we used the Newcastle-Ottawa Quality Assessment Scale to appraise study quality.[7] Both study data and data related to risk of bias were abstracted by one investigator, and were reviewed for accuracy by at least one additional investigator.

Discussions with Key Informants

We engaged experienced P4P researchers as key informants to gain insight into issues related to implementation and unintended consequences. Key informants were identified as those having expertise on pay-for-performance programs in healthcare through a review of relevant literature and through consultation with our stakeholders and Technical Expert Panel. We conducted hour-long semi-structured interviews with KIs to understand their perceptions of implementation factors that were important in both positively and negatively influencing P4P programs. Five investigators conducted independent inductive open-coding of interview notes. One investigator with qualitative research experience (KK) reviewed the investigators' codes and identified common themes.

Data Synthesis

We qualitatively synthesized the results of included studies according to an implementation framework based on the Consolidated Framework for Implementation Research (CFIR),[8] and modified for the topic in collaboration with our panel of technical experts (Fig. 1). The framework applies to P4P in healthcare generally, and describes the relationship between the features of P4P programs, external factors, implementation factors, and provider cognitive/affective and behavioral responses on processes of care and patient outcomes. This paper focuses on the relationships between implementation factors, which include implementation processes, features related to the inner and outer settings, and provider characteristics; program design features; provider cognitive/affective responses; provider behavioral responses; and the effect on processes of care and patient outcomes. Table 1 describes each category included in the framework. Due to the large number of observational trials and heterogeneity among the studies, meta-analysis was not performed.

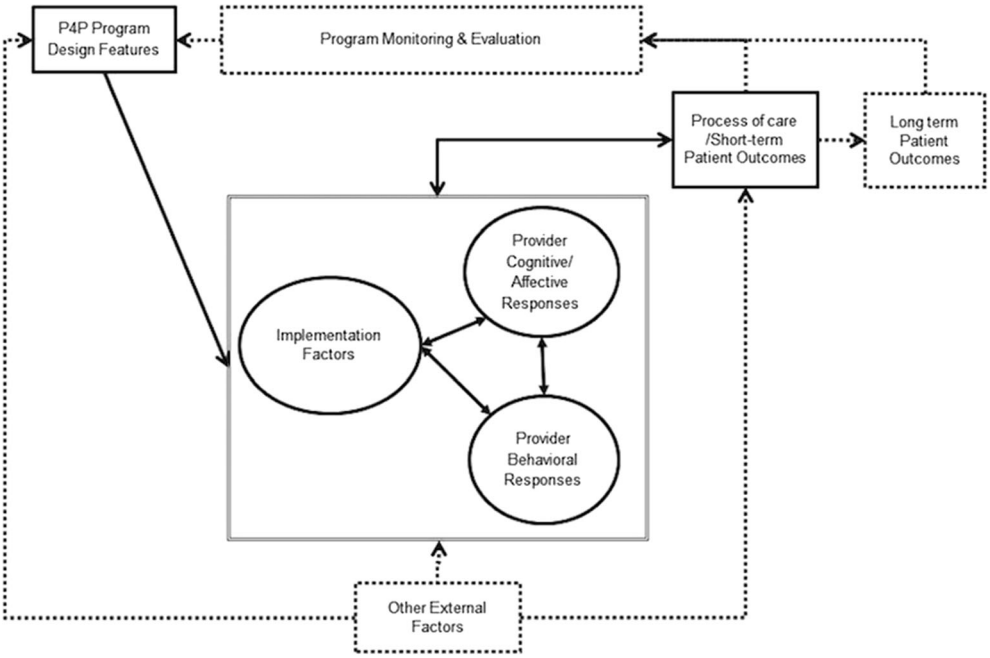

Figure 1 Conceptual framework.

RESULTS

We reviewed 1363 studies, with 509 examined at the full-text level. Forty studies met inclusion criteria, with an additional study identified by a peer reviewer, for a total of 41 (Fig. 2; see Table 2 for study characteristics; study details provided in Appendices 4 and 5, available online). Of 45 individuals invited, 14 participated in KI interviews (Appendix 6, available online).

Program Design Features (13 Studies)

We identified one prospective cohort,[9] two retrospective cohort,[10,11] and one pre-post study,[12] six cross-sectional surveys,[13–17] one economic analysis,[18] and two simulation studies.[19,20] Related to measure development, studies found that an emphasis on clinical quality and patient experience criteria was related to increased coordination of care, improved office staff interaction, and provider confidence in providing high-quality care.[11,14] Conversely, an emphasis on productivity and efficiency measures was associated with poorer provider and office staff communication.[11] In addition, one study that surveyed administrators and managers about the overall effectiveness of a P4P program found that factors predictive of the perceived effectiveness of the program included both the communication of goal alignment and the alignment of individual goals to institutional goals, while another found that providers believed that the P4P program increased a clinician's focus on

Table 1. Description of Implementation Framework Categories

Framework Category		Description
Program design features		Properties of the intervention itself such as the type of quality measure used or the size of the financial incentive
Implementation factors	Implementation processes	Actions taken to implement the P4P program such as planning, stakeholder engagement, academic detailing, audit and feedback, and whether the incentive was targeted at the team or individual level
	Outer setting	Refers to the broader health system context within which an intervention is implemented; the cultural and social norms at the state and federal level; and characteristics of the patient population
	Inner setting	Refers to characteristics of the institution or organization itself
	Provider characteristics	Refers to demographic characteristics (e.g., age, gender, race/ethnicity), as well as other factors such as experience and specialization
Provider cognitive/affective and behavioral responses		Refers to provider beliefs and attitudes. Includes cognitive response constructs such as biases, professionalism, heuristics, identification with one's organization. Also includes behavioral response constructs such as risk selection, gaming, systems improvement responses
Process-of-care and short-term patient outcomes		Includes process-of-care outcomes such as performance of recommended screening or disease monitoring, as well as patient outcomes such as achieving target disease management goals (e.g., blood pressure, cholesterol levels) and health outcomes

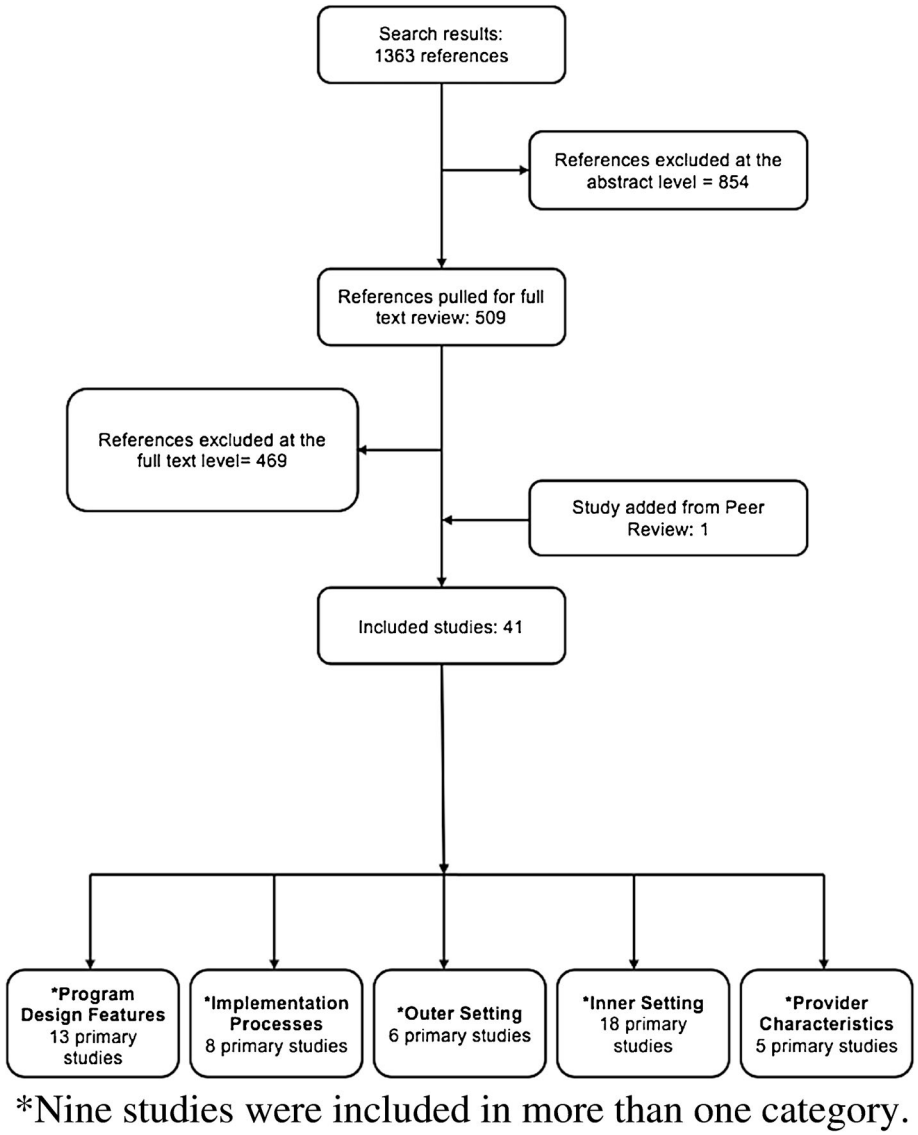

Figure 2 Literature flow.

issues related to quality of care.[12,15] Finally, one study examined different statistical methods of constructing composite measures, and found latent variable methods to be more reliable than raw sum scores.[19]

Related to incentive structures, one study examined the extent to which incentive size related to the decision to participate in P4P programs, and found that no clear amount determined decisions of whether to participate, but rather that there was a positive relationship between participation and the potential for reward.[10] Similarly, another study found that after controlling for covariates, perceived financial salience was significantly related to a high degree of performance.[13] Another study found that the underlying payment structure influenced performance, and that higher incentives may be necessary when the degree of cost sharing is lower.[9] Finally, a study examining the relationship between P4P and patient experience in California over a 3-year period found that, compared with larger incentives (>10%), smaller incentives were associated with greater improvement in provider communication and office staff interaction measures.[11] These findings were contrary to the hypotheses of the authors, who concluded that their findings may have been influenced by the tendency of practices with smaller incentives to incentivize clinical quality and patient experience measures (vs. productivity measures), which were also associated with improvements in office staff interaction.

Table 2. Study Characteristics

Included studies	*n*
Total	41
Setting:	
Ambulatory	33
Hospital	7
Nursing home	1
Country:	
United States	18
United Kingdom	17
Canada	2
Australia	1
France	1
Netherlands	1
Taiwan	1

Findings from Key Informant Interviews. Key informants 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, having too many measures increased the likelihood of negative unintended consequences. KIs also agreed that measures should reflect organizational priorities, and should 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, they suggested that improvements should be incentivized, that incentives should be large enough to provide motivation, but not so large as to encourage gaming, that penalties may be more effective than rewards, and that team-based incentives may be effective for increasing buy-in and professionalism among 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 sufficient to reinforce behavior.

Implementation Processes (8 Studies)

We identified seven cohort studies, one prospective[21] and six retrospective,[22–27] and one simulation study.[28] Three included studies[25,26,28] examined threshold changes in the QOF, and found that quality continued to increase after increases in maximum thresholds, with lower-performing providers improving significantly more than those who were performing at a high level under the previous threshold.[25,26] In addition, we identified three studies examining clinical process, and patient outcome incentives were removed from a measure. One study, of the QOF, found that the level of performance achieved prior to the incentive withdrawal was generally maintained, with some difference by indicator and disease condition.[27] Two studies examined changes in incentives within the VHA. Benzer et al. (2013) evaluated the effect of incentive removal and found that all improvements were sustained for up to 3 years.[22] Similarly, Hysong and others (2011) evaluated changes in measure status, that is, the effect on performance when measures shift from being passive monitored (i.e., no incentive) to actively monitored (i.e., incentivized), and vice versa.[23] Findings indicate that regardless of whether a measure was incentivized, all remained stable or improved over time. Quality did not deteriorate for any of the measures in which incentives were removed, and of the six measures that changed from passive to active monitoring, only two improved significantly after the change (HbA1c and colorectal cancer screening).

Findings from Key Informant Interviews. Similar to the findings reported in the literature, key informants believed that measures should be evaluated regularly (e.g., yearly) to enable continued increases in quality. Once achievement rates are high, those measures should be evaluated, with the possibility of increasing thresholds, if relevant, or replacing them with others representing areas in need of quality improvement.

KIs stressed that implementation processes should be transparent and should 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 achieve buy-in, KIs urged the engagement of stakeholders of all levels, recommended a "bottom-up" approach to program development, and strongly supported clear performance feedback to providers at regular intervals, accompanied by suggestions for and examples of how to achieve high levels of performance.

Outer Setting (6 Studies)

We identified five retrospective cohort studies[29–33] and one cross-sectional survey[17] related to the outer setting. Studies provided no clear evidence related to factors associated with region, population density, or patient population. One short-term study of the QOF reported better performance associated with a larger proportion of older patients.[33] Findings related to performance in urban compared with rural settings were inconsistent, with two studies reporting better performance by providers in rural settings,[29,32] and one finding no difference.[31]

Findings from Key Informant Interviews. Key informants discussed the importance of taking patient populations into account when designing P4P programs, stressing the importance of flexibility in larger multi-site programs to allow for targets that are realistic and that meet the needs of local patient populations.

Inner Setting (18 Studies)

We identified 15 retrospective cohort studies[30,32–45] and three cross sectional surveys[15,46,47] related to the inner setting. Studies of the QOF found that larger practices in the UK performed better in the short term,[33–35] particularly when examining total QOF points;[37] however, results varied when examining subgroups by condition or location and by indicator.[36,44,45] In addition, two studies found that group practice and training practice status was associated with higher quality of care,[33,34] while two others found no significant effect of training practice status after controlling for covariates.[35,44] Studies in the United States and other countries indicate that factors related to higher quality or greater quality improvement include culture change interventions introduced along with P4P,[46] and clinical support tools.[42] Results were mixed regarding quality improvement visits/groups and training.[15,47] Contrary to findings related to the QOF, however, differences in quality associated with P4P within independent versus group practices,[48] type of hospital (e.g., training, public, private, etc.),[30] and patient panel size/volume are less clear, with studies reporting conflicting results.[30,43]

Findings from Key Informant Interviews. KIs stressed that P4P is just one piece of an overall quality improvement program, with other important factors such as a strong infrastructure and ongoing infrastructure support (particularly with regard to information technology and electronic medical records), organizational culture around P4P and associated measures, alignment/allocation of resources with P4P

measures, and public reporting. Public reporting was described by many of our KIs as a strong motivator, particularly for hospital administrators, but also for individual providers operating within systems in which quality achievement scores are shared publically.

Provider Characteristics (5 Studies)

We identified three retrospective cohort studies[29,34,43] and two cross-sectional surveys.[13,49] Studies examining the influence of provider characteristics found no strong evidence that provider characteristics (e.g., gender, age) related to performance in P4P programs.[13,29,34,43,45]

DISCUSSION

We identified 41 studies examining factors related to the implementation of P4P programs. Studies targeted implementation features associated with the effect on process-of-care and short-term patient outcomes, as well as on provider cognitive, affective, and behavioral responses. Implementation features included those related to program design, such as factors related to the incentivized measures; implementation processes, such as updating or retiring measures; the inner and outer settings; and provider characteristics. The studies we examined 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 across health systems and organizations and the challenges related to the evaluation of complex interventions such as P4P preclude us from drawing firm 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 (Table 3). First, programs that emphasize measures targeting process-of-care or clinical outcomes that are transparently evidence-based and viewed as clinically important may inspire more positive change than programs using measures targeted to efficiency or productivity, or that 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 provider expectations for clinical quality, and there was 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 U.S. 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, and that incentivized measures must be congruent with institutional priorities, must address the needs of the institution at the local level, and must 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 should be 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, including changes to the measures incentivized and the thresholds associated with payments.

Finally, and related, 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 the notion 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 areas in greatest need of improvement.

Limitations

Our review has a number of limitations. Due to the recent report on pay-for-performance programs published by the RAND Corporation and commissioned by CMS, which focused largely on programs in the United States, and our inclusion of studies examining the UK Quality and Outcomes Framework, our review and subsequent conclusions are weighted heavily towards programs targeting ambulatory care. In addition, given the heterogeneity among P4P programs, and our goal of better understanding the important factors related to implementation, we included studies that utilized less rigorous methodology, some of which had small samples. The breadth of topics and outcomes related to implementation characteristics made it difficult to restrict our criteria by study design. Given these factors, along with the inclusion of studies examining primarily observational data, we did not formally assess strength of evidence. To better inform an understanding of implementation factors important to the success of P4P programs, we interviewed 14 key informants. As our goal was not to conduct primary research, our key informants were experienced P4P researchers in the United States and the United Kingdom. While their knowledge and experience provided us with insight into implementation processes and unintended consequences, and although they were particularly well positioned to speak to future research needs, we recognize that conversations with other stakeholders, including policymakers, program officials, hospital administrators and managers, providers and other clinical and non-clinical staff, and patients, are necessary to more fully understand the issues related to P4P.

Future Research

Despite numerous P4P programs in the United States, the United Kingdom, and elsewhere, there is a need for higher-

Table 3. Evidence and Policy Implications by Implementation Framework Category

Implementation Framework Category	Study Evidence	Themes from KI Interviews	Policy Implications
Program design features	Thirteen studies[9–20,50] 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, while 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 to 15%. • Incentives should be based on improvements, and all program participants should have the ability to earn incentives. • The magnitude of an incentive attached to a specific measure should be relative to organizational priorities. • Consider distributing incentives to clinical and non clinical staff.	• 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[21–28] examined changes in implementation, with seven specifically related to updating or retiring measures, and found: • Under both the QOF and in the VHA, removing an incentive from a measure had little impact on performance once a high performance level had been achieved. • Increasing maximum thresholds resulted in greater increases by poorer-performing practices.	• Evaluate measures regularly and consider increasing thresholds or removing incentives once high performance has been achieved. • Stakeholder involvement and provider buy-in are critical. • A bottom-up approach is effective. • Provide reliable data/feedback to providers in a non-judgmental fashion.	• P4P programs should target areas of poor performance and consider de-emphasizing areas that have achieved high performance.
Outer setting	Six studies[17,29–31,33,34,48] examined implementation factors related to the outer setting. • There is no clear evidence that setting (e.g., 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[15,30,33–48] 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 U.S. 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	Five studies[13,29,34,43,49] 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.

quality evidence to better understand whether these programs are effective in improving the quality of healthcare and the implementation factors that contribute to their success. Studies examining P4P have been largely observational and primarily retrospective, or have lacked good matched comparison groups, and research examining implementation characteristics has often been conducted with small samples. 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, with programs differing widely in their measures and incentive structures, as well as 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, including the influence of public reporting, the number and focus of measures, incentive size, structure, and target. 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, not only to examine P4P's effectiveness on processes of care and patient outcomes directly, but also to better understand and clarify the implementation characteristics important in achieving higher quality of care and in mitigating unintended consequences.

Acknowledgments: *We would like to acknowledge the contributions of our stakeholders and the Technical Expert Panel.*

Corresponding Author: *Karli K. Kondo, PhD, MA; Portland Veterans Affairs Medical Center, Evidence-based Synthesis Program, Mailcode RD71, 3710 SW U.S. Veterans Hospital Road, Portland, OR 97239, USA (e-mail: karli.kondo@va.gov).*

Compliance with Ethical Standards:

Funding: *This project was funded by the U.S. Department of Veterans Affairs, Veterans Health Administration (VHA) ESP Project #05-225.*

Prior Presentation: *The contents of this manuscript have not been presented at any conference.*

Conflict of Interest: *The authors declare that they have no conflict of interest, financial or otherwise, to disclose in relation to the content of this paper. Funding for the VA ESP is provided by the VHA Quality Enhancement Research Initiative (QUERI), Health Services Research and Development Service (HSR&D). The Department of Veterans Affairs had no role in the conduct of the study, in the collection, management, analysis, or interpretation of the data, or in the preparation of the manuscript. The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs or the U.S. government.*

REFERENCES

1. Department of Veterans Affairs. Department of Veterans Affairs Health Care Personnel Enhancement Act of 20042004.
2. United States Government Accountability Office. VA Health Care. Actions needed to improve administration of the provider performance pay and award systems: Report to congressional requesters. 2013.
3. **Kizer KW, Kirsh SR.** The double edged sword of performance measurement. J Gen Intern Med. 2012;27(4):395–7.
4. **Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S.** Does pay-for-performance improve the quality of health care? Ann Intern Med. 2006;145(4):265–72.
5. **Damberg C, Sorbero M, Lovejoy S, Martsolf G, Raaen L, Mandel D.** Measuring Success in Health Care: Value-Based Purchasing Programs. Findings from an Environmental Scan, Literature Review, and Expert Panel Discussions. Santa Monica, California. 2014.
6. **Treadwell JR, Singh S, Talati R, McPheeters ML, Reston JT.** A Framework for "Best Evidence" Approaches in Systematic Reviews. 2011.
7. **Wells G, Shea B, O'Connell D, et al.** The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed Nov 19, 2015.
8. **Damschroder L, Aron D, Keith R, Kirsh S, Alexander J, Lowery J.** Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci: IS. 2009;4(50).
9. **Kantarevic J, Kralj B.** Link between pay for performance incentives and physician payment mechanisms: evidence from the diabetes management incentive in Ontario. Health Econ. 2013;22(12):1417–39.
10. **de Brantes FS, D'Andrea BG.** Physicians respond to pay-for-performance incentives: larger incentives yield greater participation. Am J Manag Care. 2009;15(5):305–10.
11. **Rodriguez HP, von Glahn T, Elliott MN, Rogers WH, Safran DG.** The effect of performance-based financial incentives on improving patient care experiences: a statewide evaluation. J Gen Intern Med. 2009;24(12):1281–8.
12. **Young GJ, Beckman H, Baker E.** Financial incentives, professional values and performance: a study of pay-for-performance in a professional organization. J Organ Behav. 2012;33(7):964–83.
13. **Waddimba AC, Meterko M, Beckman HB, Young GJ, Burgess JF Jr.** Provider attitudes associated with adherence to evidence-based clinical guidelines in a managed care setting. Med Care Res Rev. 2010;67(1):93–116.
14. **Baek JD, Xirasagar S, Stoskopf CH, Seidman RL.** Physician-targeted financial incentives and primary care physicians' self-reported ability to provide high-quality primary care. J Prim Care Community Health. 2013;4(3):182–8.
15. **Helm C, Holladay CL, Tortorella FR.** The performance management system: applying and evaluating a pay-for-performance initiative… including commentary by Candio C. J Healthc Manag. 2007;52(1):49–63.
16. **Hadley J, Landon BE, Reschovsky JD.** Effects of compensation methods and physician group structure on physicians' perceived incentives to alter services to patients. Health Serv Res. 412006:1200+.
17. **Hearld LR, Alexander JA, Shi Y, Casalino LP.** Pay-for-performance and public reporting program participation and administrative challenges among small- and medium-sized physician practices. Med Care Res Rev: MCRR. 2014;71(3):299–312.
18. **Walker S, Mason AR, Claxton K, et al.** Value for money and the quality and outcomes framework in primary care in the UK NHS. Br J Gen Pract: J R Coll Gen Pract. 2010;60(574):e213–20.
19. **Chen T-T, Lai M-S, Lin IC, Chung K-P.** Exploring and comparing the characteristics of nonlatent and latent composite scores: implications for pay-for-performance incentive design. Med Decis Mak. 2012;32(1):132–44.
20. **Werner RM, Dudley RA.** Making the 'pay'matter in pay-for-performance: implications for payment strategies. Health Aff. 2009;28(5):1498–508.
21. **Andriole KP, Prevedello LM, Dufault A, et al.** Augmenting the impact of technology adoption with financial incentive to improve radiology report signature times. J Am Coll Radiol. 2010;7(3):198–204.
22. **Benzer JK, Young GJ, Burgess JF Jr, et al.** Sustainability of quality improvement following removal of pay-for-performance incentives. J Gen Intern Med. 2013;29(1):127–32.
23. **Hysong SJ, Khan MM, Petersen LA.** Passive monitoring versus active assessment of clinical performance: impact on measured quality of care. Med Care. 2011;49(10):883–90.
24. **Shih T, Nicholas LH, Thumma JR, Birkmeyer JD, Dimick JB.** Does pay-for-performance improve surgical outcomes? An evaluation of phase 2 of the premier hospital quality incentive demonstration. Ann Surg. 2014;259(4):677–81.
25. **Feng Y, Ma A, Farrar S, Sutton M.** The tougher the better: an economic analysis of increased payment thresholds on performance of general practices. Health Econ. 2014.
26. **Kontopantelis E, Doran T, Gravelle H, Goudie R, Siciliani L, Sutton M.** Family doctor responses to changes in incentives for influenza immunization under the UK quality and outcomes framework pay-for-performance scheme. Health Serv Res. 2012;47(3 Pt 1):1117–36.

27. **Kontopantelis E, Springate D, Reeves D, Ashcroft DM, Valderas JM, Doran T.** Withdrawing performance indicators: retrospective analysis of general practice performance under UK quality and outcomes framework. BMJ (Clin Res ed). 2014;348:g330.
28. **Caley M, Burn S, Marshall T, Rouse A.** Increasing the QOF upper payment threshold in general practices in England: impact of implementing government proposals. Br J Gen Pract: J R Coll Gen Pract. 2014;64(618):e54–9.
29. **Arrowsmith ME, Majeed A, Lee JT, Saxena S.** Impact of pay for performance on prescribing of long-acting reversible contraception in primary care: an interrupted time series study. PLoS One. 2014;9(4):e92205.
30. **Bhattacharyya T, Mehta P, Freiberg AA.** Hospital characteristics associated with success in a pay-for-performance program in orthopaedic surgery. J Bone Joint Surg Am Vol. 2008;90A(6):1240–3.
31. **Greene J.** An examination of pay-for-performance in general practice in Australia. Health Serv Res. 2013;48(4):1415–32.
32. **Kirschner K, Braspenning J, Jacobs JE, Grol R.** Experiences of general practices with a participatory pay-for-performance program: a qualitative study in primary care. Aust J Prim Health. 2013;19(2):102–6.
33. **Ashworth M, Armstrong D.** The relationship between general practice characteristics and quality of care: a national survey of quality indicators used in the UK quality and outcomes framework, 2004–5. BMC Fam Pract. 2006;7:68.
34. **Ashworth M, Schofield P, Seed P, Durbaba S, Kordowicz M, Jones R.** Identifying poorly performing general practices in England: a longitudinal study using data from the quality and outcomes framework. J Health Serv Res Policy. 2011;16(1):21–7.
35. **Walker N, Bankart J, Brunskill N, Baker R.** Which factors are associated with higher rates of chronic kidney disease recording in primary care? A cross-sectional survey of GP practices. Br J Gen Pract: J R Coll Gen Pract. 2011;61(584):203–5.
36. **Wang Y, O'Donnell CA, Mackay DF, Watt GC.** Practice size and quality attainment under the new GMS contract: a cross-sectional analysis. Br J Gen Pract: J R Coll Gen Pract. 2006;56(532):830–5.
37. **Doran T, Campbell S, Fullwood C, Kontopantelis E, Roland M.** Performance of small general practices under the UK's quality and outcomes framework. Br J Gen Pract. 2010;60(578):335–44.
38. **Morgan CL, Beerstecher HJ.** Primary care funding, contract status, and outcomes: an observational study. Br J Gen Pract: J R Coll Gen Pract. 2006;56(532):825–9.
39. **Tahrani AA, McCarthy M, Godson J, et al.** Impact of practice size on delivery of diabetes care before and after the quality and outcomes framework implementation. Br J Gen Pract: J R Coll Gen Pract. 2008;58(553):576–9.
40. **Gemmell I, Campbell S, Hann M, Sibbald B.** Assessing workload in general practice in England before and after the introduction of the pay-for-performance contract. J Adv Nurs. 2009;65(3):509–15.
41. **Dalton ARH, Alshamsan R, Majeed A, Millett C.** Exclusion of patients from quality measurement of diabetes care in the UK pay-for-performance programme. Diabet Med. 2011;28(5):525–31.
42. **Kruse GR, Chang Y, Kelley JH, Linder JA, Einbinder JS, Rigotti NA.** Healthcare system effects of pay-for-performance for smoking status documentation. Am J Manage Care. 2013;19(7):554–61.
43. **Li J, Hurley J, Decicca P, Buckley G.** Physician response to pay-for-performance: evidence from a natural experiment. Health Econ. 2013.
44. **Norbury M, Fawkes N, Guthrie B.** Impact of the GP contract on inequalities associated with influenza immunisation: a retrospective population-database analysis. Br J Gen Pract: J R Coll Gen Pract. 2011;61(588):e379–85.
45. **Vamos EP, Pape UJ, Bottle A, et al.** Association of practice size and pay-for-performance incentives with the quality of diabetes management in primary care. CMAJ: Can Med Assoc J. 2011;183(12):E809–16.
46. **Miller SC, Looze J, Shield R, et al.** Culture change practice in u.s. Nursing homes: prevalence and variation by state medicaid reimbursement policies. The Gerontologist. 2014;54(3):434–45.
47. **Begum R, Smith Ryan M, Winther CH, et al.** Small practices' experience with EHR, quality measurement, and incentives. Am J Manage Care. 2013;19(10 Spec No):eSP12–8.
48. **Kirschner K, Braspenning J, Akkermans RP, Jacobs JE, Grol R.** Assessment of a pay-for-performance program in primary care designed by target users. Fam Pract. 2013;30(2):161–71.
49. **Saint-Lary O, Bernard E, Sicsic J, Plu I, Francois-Purssell I, Franc C.** Why did most French GPs choose not to join the voluntary national pay-for-performance program? PLoS One. 2013;8(9):e72684.
50. **Torchiana DF, Colton DG, Rao SK, Lenz SK, Meyer GS, Ferris TG.** Massachusetts general physicians organization's quality incentive program produces encouraging results. Health Aff. 2013;32(10):1748–56.