Robert Wood Johnson Foundation

THE SYNTHESIS PROJECT

NEW INSIGHTS FROM RESEARCH RESULTS

RESEARCH SYNTHESIS REPORT NO. 13
DECEMBER 2007

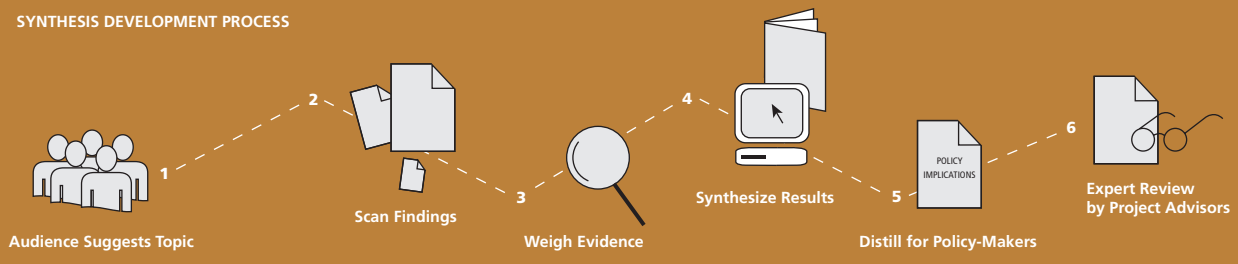
Jon B. Christianson, Ph.D.,
Sheila Leatherman, M.S.W,
and Kim Sutherland, Ph.D.

Paying for quality: Understanding and assessing physician pay-for-performance initiatives

See companion Policy Brief available at *www.policysynthesis.org*

TABLE OF CONTENTS

THE SYNTHESIS PROJECT (Synthesis) is an initiative of the Robert Wood Johnson Foundation to produce relevant, concise, and thought-provoking briefs and reports on today's important health policy issues. By synthesizing what is known, while weighing the strength of findings and exposing gaps in knowledge, Synthesis products give decision-makers reliable information and new insights to inform complex policy decisions. For more information about the Synthesis Project, visit the Synthesis Project's Web site at *www.policysynthesis.org*. For additional copies of Synthesis products, please go to the Project's Web site or send an e-mail request to *pubsrequest@rwjf.org*.



Introduction

There is now considerable interest among private and public health care purchasers in using financial incentives to improve the quality of care delivered by physicians, as well as some disagreement over the likely consequences (47, 56). The concept has gained traction as a way for purchasers to better align physician payment and quality of care delivered.

Pay-for-performance initiatives (P4P)[1] are being pursued by state Medicaid programs and are of great interest to Medicare. Recently, Medicare linked its 2007 payment upgrades for physicians to the reporting of performance data, a step some regard as laying the foundation for a P4P program (41) as recommended by the Institute of Medicine (40). Meanwhile, a 2006 survey reported that 28 states have adopted some type of pay-for-performance initiative in their Medicaid programs (43), and that half of these initiatives have been in existence for five years or more. It is not entirely clear how many programs are directed at physicians, but it would appear that most contain at least a physician component.

Past experience—including with managed care—shows that financial incentives can be a powerful driver for physician behavior.

To date, however, policy-makers have had little information on the effectiveness of P4P initiatives in shifting physician practice. They are interested in knowing to what extent and under what circumstances P4P will improve the quality of care delivered by physicians. This synthesis report reviews the available evidence on this issue, addressing five questions:

1. What explains the current widespread interest in physician P4P?
2. How are current incentive programs structured and how prevalent are they?
3. What performance measurement issues does physician P4P raise?
4. How do physicians perceive quality incentive programs?
5. What is the research evidence on the impact of P4P?

These are important questions for federal and state policy-makers who have implemented, or are moving towards implementing, P4P initiatives in Medicare and Medicaid, as well as for large purchasers who seek to do the same. The findings will assist policy-makers and purchasers in clarifying expectations regarding P4P and implementing it effectively.

What explains the current widespread interest in physician P4P?

Linking physician financial incentives to quality performance metrics is not new (52). During the early 1990s, HMO physician payment methodologies featured a complex blend of incentives mostly designed to constrain service use and encourage the delivery of care in lower-cost settings and by less expensive providers, raising concerns about quality of care. Even then, however, many of these arrangements included quality-related incentives. For example, in a survey of HMOs conducted in 1992, about 20 percent of responding organizations said their payments to physicians incorporated some reimbursement for performance on quality of care, with 20 percent also reporting physician payments tied to consumer satisfaction measures (49).

1 The idea has been discussed under the general rubrics of "Quality-Based Purchasing" (22) or "Value-Based Purchasing" (62, 64), but the term "pay-for-performance" (abbreviated to P4P) has become increasingly popular as a descriptive label.

Introduction

During the early 1990s, a group of HMOs and large employers agreed on a set of performance measures–referred to as "HEDIS"[2]–to be reported annually on a voluntary basis by HMOs (8) (see Appendix II). It was expected that employers would consider the HEDIS scores of HMOs when making their health benefits contracting decisions. The development of the measures focused attention on issues relating to quality measurement, encouraged health plans to work with contracting physicians on initiatives to improve quality, created momentum for the adoption of electronic medical records systems in hospitals and physician offices and generated a set of measures that were widely accepted as legitimate quality indicators, at a population level. Nevertheless, HEDIS measures were largely limited to a small number of chronic diseases and the provision of a limited number of preventive services.

The HEDIS initiative led to other performance reporting initiatives. The dissemination of data on HEDIS measures spawned numerous efforts to refine quality measures and to develop measures that could be used to assess the performance of hospitals, physicians and physician groups. Employers came to believe that measures constructed at the provider level would be more useful to employees in making their health care decisions than measures of overall HMO performance. With most community providers available in most PPO networks, employees were increasingly choosing among providers, not plans (6, 7, 14, 31).

During this time, research evidence was accumulating that the quality of ambulatory care left much to be desired. There was considerable opportunity for improvements in the quality of care delivered in physician offices (46) and a growing consensus that several different approaches to accomplishing quality improvement were needed (34, 35). In 2001, the Institute of Medicine (IOM) report on "Crossing the Quality Chasm" galvanized purchasers and physician organizations around the challenge of improving quality (39). A key recommendation of that report, and a subsequent IOM report (40), was that payment incentives for providers needed to be "realigned" to support quality improvement.

Realigning payment incentives in the 1990s was part of a larger strategy to contain costs being developed by large employers and their benefit consultants in response to double-digit increases in employer health care costs that began in the late 1990s (31). Under this strategy, employees would share more health care costs at the point of service, creating financial incentives for them to play a more substantial role in health care decisions, including choice of provider. Employers and health plans would increase their efforts to measure provider performance and to disseminate information on the cost and quality of providers to employees and plan enrollees, resulting in a more "market-driven" system. The reward to higher quality physicians presumably would be that, over time, more patients would seek out their services, increasing practice revenues. Recognizing, however, that this "market reward" could take some time to develop, some purchasers also saw value in implementing more direct rewards for better quality care (71).

2 Healthcare Effectiveness Data and Information Set, see Appendix II.

Introduction

The present interest in reforming provider payment to reward quality of care is the result of a variety of forces:

- The evolution of quality measures and experience gained in applying those measures to health plans
- Research suggesting significant opportunities for quality improvement
- Endorsement of P4P from high-profile national bodies, including the Institute of Medicine
- The support of some large purchasers, who see it as a potentially valuable complement to their "consumer-oriented" strategies

In addition to these forces, policy analysts have argued that overall payment approaches for physicians are deficient in many ways, including their impact on quality (see Figure 3 for descriptions of common payment approaches) (16, 54).

Findings

How are current incentive programs structured and how prevalent are they?

How prevalent are quality-related incentive programs for physicians?

Many, and perhaps the majority of, health plans now have P4P programs, but it is difficult to assess their growth and impact over time. Survey results tell us something about the presence of P4P programs (Figure 1), but it is not clear if their number or the share of physicians affected has increased substantially over time. Efforts to track the development of physician P4P programs are hampered by the lack of an annual survey conducted in a systematic way that generates publicly available results.

Discussions of these programs typically highlight a relatively small number of ambitious efforts (e.g., Integrated Healthcare Association (IHA), Bridges to Excellence and the U.K. initiative; see Appendix III). Further, while many large health plans have P4P programs, these programs may apply to only a subset of contracting physician practices and individual physicians are not always aware of the P4P incentives in their contracts.

Figure 1. Prevalence and features of P4P programs

Author	Data used	Percent of plans with P4P initiatives	Percent of identified P4P initiatives rewarding physician performance
Rosenthal (58)	Author identified 37 P4P programs		76
Rosenthal (61)	Systematic survey of health plans in 40 randomly selected health care markets	52	90
Trude (70)	Survey of health plans in 12 Community Tracking Study communities	77*	

* Under consideration, in planning stage or executed.

Just over one-quarter of primary care physicians report having quality-based compensation incentives. Using data from four physician surveys, Reschovsky and Hadley (53) found that in 2004/2005 just over one-quarter (28 percent) of primary care physicians in group practices reported quality-based incentives in their compensation arrangements, modestly higher than the share reporting such incentives in 1996/1997 (26 percent). The upward trend is partly due to physician movement to larger practice settings more likely to have quality-based incentives.

The most common quality incentives facing U.S. primary care physicians are for meeting specific clinical targets and for patient satisfaction (Figure 2). Incentives for meeting clinical targets are encountered by 23 percent of U.S. primary care physicians, while those for patient satisfaction and quality of care processes are encountered by 20 and 19 percent of primary care physicians respectively.

Figure 2. Percent of U.S. primary care physicians facing specific financial incentives for quality, 2006

Financial incentive for quality	Percent receiving or having potential to receive this incentive
Achieving certain clinical targets	23
High ratings for patient satisfaction	20
Participating in quality improvement activities	19
Enhanced preventive care activities	12
Managing patients with chronic disease/complex needs	8

Source: 2006 Commonwealth Fund International Health Policy Survey of Primary Care Physicians

How are existing incentive programs structured?

A large number of authors have addressed the great number and variety of decisions that must be made in designing physician incentives to reward quality, with the most comprehensive description supplied by Dudley and Rosenthal (23) in their "Decision Guide to Purchasers". The most important design decisions concern the type and size of incentives and the choice of measures to assess performance.

New incentive programs are layered on top of existing payment arrangements and, to some degree, seek to counter their incentives (Figure 3). It is possible that physician incentive programs, identical in other respects, could have different impacts on physician behavior depending on existing payment incentives. In fact, the situation for most physicians is likely to be quite complex. Within any single incentive program, there are typically multiple different performance measures, with each having a different potential for generating financial rewards. For a physician receiving payments from a variety of health plans and government programs, all with somewhat different basic payment approaches and different incentive schemes to reward quality, making rational decisions about how to allocate time and practice resources is likely to be daunting.

As members of physician groups, many physicians are insulated from the direct effects of P4P incentives. These physicians are typically paid using some combination of a base salary and a "productivity incentive;" that is, a reward connected to the number of patients seen or the amount of practice revenue generated. Practices aggregate revenues from all payers, including payments from P4P programs, and distribute these revenues according to a formula approved by the physicians in the group. In this situation, the direct connection between the financial incentives of any single P4P program and the practice behavior of physicians would be mitigated by group decisions about contracts with payers, and by policies relating to the distribution of practice income.

Findings

Figure 3. Physician payment approaches and the incentives they create

Approach	Description	Potential incentives created
Capitation	The physician agrees to deliver a specified list of health services for a fixed amount per person. The physician bears financial risk.	• The physician might act too aggressively in constraining service use, eliminating some "necessary" as well as some "unnecessary" services. The result could be lower quality of care for patients especially if there is no sharing of risks or surpluses, if the capitated contract is short-term in nature and if contract renewal does not depend on measures other than costs. • Conversely, if physician organizations reimbursed by capitation payments care for an enrolled population over a period of time, they have an incentive to provide services that maintain or improve the health of that population, as this will be financially beneficial in the long term.
Fee-for-service	Physicians are paid for each unit of service provided.	• This form of payment contains a powerful incentive for "over-provision" of services and necessitates a substantial amount of costly monitoring on the part of the payer. • There is a risk to patient health associated with "over-treatment," just as there is with "under-treatment" (39). • To modify the incentives under fee-for-service, arrangements based on payment per episode, payment per admission or evidence-based case rates have been introduced (19). These and similar approaches bundle services for payment purposes, creating incentives for physicians to limit the services they provide in response to a specific event. However, unlike capitation, physicians receive more revenues the greater number of events they treat.
Salary	The physician is paid a fixed amount per time period.	• There is no incentive to deliver unnecessary services, nor is there an incentive for "under-provision," except to the degree that physicians may "shirk" under salaried arrangements. • There is no particular incentive under a pure salary method of payment for physicians to deliver high quality care, so there typically is a heavy reliance on enforcement of rules and procedures thought to enhance quality. • The result could be quality enhancing or, to the degree that rule enforcement limits physician ability to bring professional judgment to bear in treatment decisions, result in lower quality of care.
Budgets	Physicians may be reimbursed through a negotiated budget process at the organizational level. This method of payment is most often observed internationally in government administered health care systems.	• The nature of the incentives in this payment arrangement can resemble capitation, when the number of individuals served in a given period is relatively fixed, and the organization is at risk for budget over-runs and can keep savings. • Or, the incentives can resemble those of salaried physicians when the organization serves patients who seek care, but does not assume responsibility to provide care to a fixed number, or enrolled group, of individuals for a specified time period.

Again, different physician groups are likely to respond to the same quality-related incentive scheme in different ways, depending on their cost structures, infrastructure supporting care delivery and general culture (69). But, in any case, the link between the quality incentive provided by a payer and the response of an individual physician in treating a specific patient is likely to be indirect at best.

A key decision point in developing P4P programs is whether incentives should reward improvement or meeting benchmarks (Figure 4). The different payment methods actually used by P4P programs have been described in a number of publications (23, 24, 52, 57, 58, 61, 63, 72, 73). The argument in favor of rewarding improvement is that there are potentially greater gains to be made in the quality of care received by patients of low-performing physicians. If rewards are made only for achieving benchmarks, and these benchmarks are set at a high level, low-performing physicians may be discouraged from making the effort to improve their quality. The arguments in favor of rewarding achievement of target levels of performance are essentially the reverse. That is, physicians who deliver superior care deserve to be rewarded for their efforts. Rewarding low-performing physicians could create adverse incentives for high-performing groups and raise questions about the credibility of the payer's efforts.

The problem of whether to pay for improvement or for achieving care benchmarks was underscored by the experience of Pacificare's incentive program, which used quality benchmarks to allocate reward dollars (see Appendix III). Rosenthal et al. (60) reported that physician groups that had high scores on quality indicators at the program's inception and showed little subsequent improvement received the bulk of the reward dollars distributed through this program.

A second important design decision is whether performance targets should be fixed or relative. The use of fixed quality benchmarks, as in the PacifiCare program (Appendix III), means that all physicians who meet the benchmarks are rewarded. From a physician's point of view, this "certainty" is desirable, as the practice can weigh the costs of making the changes necessary to achieve the benchmark against a certain reward. This could encourage physician practices to invest in quality improvement activities, depending on the size of the reward (73). However, from a payer's point of view, the total amount that will be spent under the P4P program becomes uncertain, unless it is capped as part of the program design. If the benchmarks are set relatively low, in an attempt to encourage physician efforts, the cost of the program could exceed payer expectations. This apparently was the case in the first year of the U.K.'s physician pay-for-performance program (20, 30; Appendix III). Benchmarks were set at a 75 percent adherence level, but average performance was at the 93 percent level, resulting in much larger than expected government payments and a general consensus that targets were set "too low" (30).

An alternative for payers is to reward physician practices for being in the top "tier" of physicians eligible to receive awards (e.g., the top ten percent of practices). In effect, physician practices compete against each other for a fixed amount of reward money, making the program easier for payers to budget (23). For physicians, however, the relationship between their performance and the probability of receiving a reward depends not only on their efforts, but also on the efforts of other physician practices. A physician practice could show great improvement, and even exceed national performance benchmarks, but not be rewarded if other practices do better still. Uncertainty regarding receipt of the reward could discourage physician practices from investing in the infrastructure changes necessary to improve the quality of their care.

Figure 4. Decision points in designing P4P incentives

		What type of targets?	
		Fixed • *Provides certainty for physicians* • *Payers are uncertain of costs*	**Relative** • *Less physician control* • *Payers may have more certainty*
What is rewarded?	**Improvement** • *Low achievers have stronger incentives to improve quality, but high achievers are "punished"* • *Rewards may go to physicians whose performance does not meet quality standards*	**Example:** Rewards physicians with X percent improvement on mammogram rate	**Example:** Rewards physicians with mammogram rate improvement in top X percent
	Achieving Benchmarks • *Rewards superior physicians, but without motivating improvement* • *Incentives may be out of reach for low performers*	**Example:** Rewards physicians with X mammogram rate	**Example:** Rewards physicians with mammogram rate in top X percent

The frequency with which HMOs choose each of these approaches is reported by Rosenthal et al. (61). (There are no systematic data regarding their use by other payers, such as PPOs.) Of 113 HMOs responding to a survey and reporting a physician incentive program, 20 percent said they paid for improvements in physician performance and 62 percent paid for achievement of a fixed performance threshold. Forty percent of HMOs said the average payment was less than five percent of their total payment to physicians, while 28 percent said that the maximum bonus a physician could receive was less than five percent.

What performance measurement issues does physician P4P raise?

Payers typically employ a mix of performance measures in their physician incentive programs, including measures of clinical care, patient satisfaction, use of information technology, patient satisfaction scores and indicators of practice efficiency. The clinical measures are used as direct indicators of quality of care. Typically, they relate to diabetes care, blood pressure control, asthma, anti-depressant medications, cholesterol management, screening tests and immunizations for children.

Clinical performance measures in most P4P programs draw heavily from HEDIS (57; Appendix II). This takes advantage of the fact that health plans and physicians have experience collecting and reporting data on these measures, costs for these activities are relatively low (many of the measures are constructed using existing claims files maintained by the plans) and the measures are familiar to employers. However, as discussed below, payers face several measurement challenges when they attempt to construct HEDIS and related performance indicators at the physician or physician practice, as opposed to the health plan, level (31). In addition, HEDIS measures address mostly processes of care and not health care outcomes, and only target recommended care for certain conditions.

Findings

How is risk adjustment carried out?

Physicians who attract more than their share of clinically complicated patients may find it difficult to score well on quality indicators that are based on patient outcomes. Or, the recommended clinical processes embodied in the performance indicators may not be appropriate for patients with multiple, complicated conditions. When this is the case, it seems fair to "adjust" physician scores to reflect differences in "patient mix", but risk-adjustment methods may not be applicable to a pay-for-performance program, or they may not be acceptable to physicians (47).

A straightforward way to address this problem is to allow physicians to exclude patients from performance measurement who have certain pre-specified characteristics. This form of "risk adjustment", adopted in the U.K.'s pay-for-performance program (Appendix III), seemed to foster a certain amount of "gaming". Physicians who excluded larger percentages of their patients from performance measurement achieved higher performance scores (20). Also, when performance is measured at the individual physician level, risk adjustment by exclusion creates the possibility that too few patients of particular types will be left in the practice to reliably measure physician performance.

What is an adequate sample for measuring performance?

Substantial variation in physician performance metrics from year-to-year based on random effects can challenge the credibility of pay-for-performance programs. This can happen when the number of patients in a physician's practice with a particular clinical problem (e.g., diabetes) is relatively small or when the number of patients associated with the payer implementing the pay-for-performance program is small. As the number of patients used to calculate performance increases, the impact of random effects on the measures is more likely to be "averaged out," and changes over time and across physicians are more likely to accurately reflect differences in performance. Nonetheless, some research suggests that, for certain types of patients, it may be difficult to construct reliable measures of performance at the individual physician level (38).

Several steps have been taken to address this problem, including restricting performance measures to care provided for patients with very common conditions, measuring performance at the physician group, rather than the individual physician, level[3] and aggregating data across payers when constructing performance measures.[4]

What is the justification for using claims data?

The use of claims data to construct physician performance measures is attractive because the data are readily available and their use minimizes data collection and reporting costs for physicians. But several issues are associated with this approach: there are a limited number of measures that can be constructed from these data; physicians raise questions about the accuracy of claims data for measuring their performance; and there is uncertainty about how sensitive claims-based measures are to changes in performance. The increased use of electronic medical records by physicians could alleviate some of these concerns, but creates its own set of complications relating to compatibility across physician offices and payers.

3 Only 13 percent of responding HMOs in Rosenthal et al. (61) focused incentives on individual physicians.
4 The IHA initiative aggregates data across multiple payers, focusing on performance at the medical group level (Appendix III).

How can measurement address the issue of multiple providers?

Not all patients have an easily identifiable "medical home". The treatment of patients with chronic conditions typically involves multiple practitioners, including several physicians (55). This raises the issue of how to connect patients to physicians for the purpose of performance measurement, especially in PPO benefit structures. Claims-based algorithms have been developed for this purpose, but physicians who are "assigned" patients under these algorithms do not necessarily see this process as fair because both the receipt and the amount of the reward can be affected by the actions of physicians who they may not know and who are not connected with their practices.

How many measures should be tracked?

The use of a limited number of performance measures in pay-for-performance programs has advantages. It can focus attention on areas with the greatest potential for quality improvement and, by concentrating incentive payments on these areas, capture the attention of physicians. However, directing attention to a few areas of care could divert resources away from treating patients with other conditions. As a result, quality could improve in the areas being rewarded, but decline for other diseases and conditions.

What is the impact of patient compliance?

Measures of patient outcomes are affected by the decisions of patients as well as the actions of physicians. This raises an issue of fairness, as different physicians treat patients with different levels of knowledge and financial resources. There is concern that physicians who treat lower income, less educated patients may perform relatively poorly on some measures because their patients are less able to effectively manage their conditions or face financial barriers in accessing care. For example, lower income women may be less likely to seek mammograms because of the cost of transportation to the physician's office or due to a lack of health insurance, and patients with low levels of "health literacy" may not fully understand the instructions for chronic care medication. Practices serving predominately less educated, lower income patients may struggle to generate adequate revenues under existing payment systems, and pay-for-performance programs may provide them with little opportunity to increase their revenues. In fact, concern was expressed prior to implementation of the U.K.'s pay-for-performance program that physician practices located in low-income areas might, over time, relocate to more affluent areas in order to improve their scores on performance indicators (55).

How do physicians perceive quality incentive programs?

Research suggests that efforts to improve clinical care processes seldom succeed without physician support and engagement. Several authors have made the same point with respect to pay-for-performance (12, 67) and physician engagement is deemed essential in the AMA's Guidelines for Pay-for-Performance Programs. The receptivity of physicians to financial incentive programs that reward quality could well be a critical factor in determining their success.

There have been several published and unpublished studies that have explored the views of physicians and practice administrators about the use of financial incentives to reward quality. These studies collect data through in-depth interviews of small numbers of physicians (9, 67) and practice administrators (4, 10) and through physician surveys (12, 74). Their results generally suggest that:

- Physicians support the concept of financial incentives that reward quality (10, 12, 67).
- Physicians have little confidence in the ability of payers, and specifically health plans, to design and carry out reward systems that are fair and effective (10, 12, 67, 74).
- Physicians who are delivering care under a P4P program may not know about the program or how it works (9).
- Physicians are concerned about the possible proliferation of P4P programs associated with different payers, and the costs this could impose on their practices.
- Physicians perceive that there is a risk of unintended consequences resulting from physician P4P (12).

What is the research evidence on the impact of P4P?

Assessing the impact of financial incentives that reward physician performance on quality measures is complex. In this section, we review the research findings on three questions:

- Does P4P result in better quality care?
- Does P4P result in other, intended or unintended, changes in physician practices?
- Do the financial benefits of P4P outweigh the costs?

Does P4P result in better quality care?

With Medicare and Medicaid at various stages of designing and implementing programs, understanding whether P4P results in better quality is a critical policy question. There are two types of research studies with findings relevant to this question: controlled experiments and program evaluations. (For brief summaries of specific experiments and programs, see Appendices V and VI.)

Controlled Experiments in the United States

These studies typically involve relatively small numbers of physician practices and patients. The practices are randomly assigned to a group exposed to incentive payments and a group that is not. Data are collected on a very limited number of quality measures, typically screening procedures or immunizations, before and after the incentive payments are put in place. The expectation is that, at the end of the study period, the physicians practicing under the incentive plan will show greater improvement with respect to the chosen quality measures than the other practices.

The strongest controlled studies provide little evidence that financial incentives improved quality of care. The results may have been due to the structure of the programs, the small payments or the difficulty untangling the impact of incentives from other quality improvement approaches.

Eight different review articles have been published that address, at least in part, controlled experiments and their findings. Six of these reviews took a broad approach in searching the literature (4, 16, 22, 51, 59), while one limited its attention to preventive care (68) and another to immunizations (1). In practice, the search strategy employed made little difference, as virtually all of the incentive schemes involved paying for preventive care of some type.

All of the reviews incorporated a core group of studies: Grady et al. (32); Kouides et al. (42); Hillman et al. (36, 37), and Fairbrother et al. (25, 26). In general, the review articles concluded that these controlled studies provided very little evidence that financial incentives improved the quality of care provided by physicians. The authors offered several explanations for why the controlled experiments were not more effective in improving quality of care, including:

- Substantial quality improvements sometimes were observed in the comparison group of physicians, making it difficult for the "incentive group" to demonstrate even greater improvement.
- The payments were relatively small in some cases and/or constituted a small portion of total practice revenues; as a result, the incentives may have been too weak to motivate physicians to respond, especially given that the experiments were time-limited.
- In some studies, it was difficult to separate the effects of the financial incentives from other concurrent efforts by practices to improve quality.
- There were weaknesses in implementation, especially in communication with physicians participating in the study.
- Improvement on preventive care measures depends on the actions of patients as well as physicians. In studies where physician practices served economically disadvantaged patients, financial and other patient-related barriers to care may have limited the ability of physicians to increase use of preventive services on the part of their patients.

In a study where significant improvements on performance measures were reported, the authors found that these improvements primarily reflected the better documentation of care (25). They conducted a subsequent study to determine if, over time, actual quality of care improvements would occur, but again concluded that much of the improvement observed in the extended study was due to better documentation. Hillman et al. (36, 37) concluded that there was no evidence that financial incentives improved care in either of their studies. Through further analysis, they found that only about half of the physician practices in the intervention group knew about the financial incentives, despite efforts to communicate with them about the program.

Policy-makers involved in "real world" initiatives are likely to have limited interest in the results of controlled experiments. One reason is that controlled experiments are mostly designed to be time-limited research projects. Under these circumstances, physicians may not find it financially or professionally attractive to invest in the practice changes necessary to improve their scores on performance measures. Larger scale and fully implemented P4P programs are likely to be perceived as permanent by physicians and, possibly, as harbingers of future changes in reimbursement policies across all payers. The effect of the same set of incentives could be quite different in these two types of programs. A second issue relating to controlled experiments is their small scale. Even if study findings supported the use of financial incentives to improve quality, it might not be possible to "scale up" the study design in a real world setting. Finally, controlled experiments typically use a very limited number of performance measures, and these measures may, or may not, be the quality measures of interest to public sector incentive programs for physicians. (For a general discussion of the strengths and limitations of different research approaches used in assessing the impact of financial incentives intended to improve the quality of physician care, see Appendix IV.)

Findings

Program Evaluations in the United States

Because they assess the "real world" application of interventions, program evaluations produce results that are the most relevant to policy-makers. These evaluations address incentive programs that have been implemented by large purchasers, primarily health plans (2, 5, 15, 27, 33, 44, 45, 48, 60). There are no review articles that synthesize their findings.

Program evaluations of P4P initiatives show more positive results than findings from controlled experiments (Appendices V and VI). Every program evaluation found improvement in one or more quality indicator. The evaluations also provide useful insights into measurement and implementation issues. As one would expect, however, given the "real world" settings in which the financial incentives were implemented, the impact of incentives themselves on quality improvement cannot be determined with complete confidence.

- Most P4P programs combine financial incentives with other efforts to improve care, suggesting that program implementers typically view P4P as one part of a multi-faceted organizational strategy supporting quality improvement. From an evaluation perspective, this makes it extremely difficult to assess the incremental contribution of P4P to any observed quality improvements.
- Most studies have not used contemporaneous control groups, raising the possibility that observed improvements might have occurred without the program.
- Most program evaluations have focused on a subset of program quality indicators, so it is not possible to assess how P4P has affected quality of care broadly defined.
- Physician participation in P4P programs typically has been voluntary, raising the question of whether the subset of physicians observed in evaluations consists primarily of those who expected to score well under program specifications. If this is the case, generalizing evaluation findings to all physicians may not be warranted.
- Most evaluations have focused exclusively, or to a significant degree, on diabetes care. Their results may not generalize to other types of care.

Evaluations of additional P4P programs (including the IHA effort (18)) soon will be forthcoming, and they promise to add substantially to the present knowledge base. However, some of the challenges that confronted the early evaluations will continue. Consequently, it may not be possible to obtain a definitive estimate of the impact of financial incentives alone on physician performance.

Does P4P result in other, intended or unintended, changes in physician practices?

To date, there is little evidence on the secondary effects of P4P initiatives. This is not surprising given the small number of published P4P program evaluations and the fact that these effects might only develop over time. Despite the lack of hard evidence, the literature on P4P speculates on a number of possible changes that could result from implementing financial incentives rewarding quality of physician care, many of which are viewed as negative. For instance, Roland (55) identified possible undesirable "secondary effects" that concerned implementers of the National Health Service (NHS) P4P program. These included: physicians may move their practices to areas where they believe patients can more effectively manage their own care; coordination of care could decline, especially for patients with multiple illnesses; physicians might focus on improving care only in areas addressed by financial rewards; and practice administrative costs could increase. Casalino and Elster (13) also expressed concern that P4P programs for

physicians in the United States could affect the care received by minorities in an adverse manner, exacerbating existing racial and ethnic disparities in care. Rosenthal (56) has suggested that, while secondary effects such as these cannot be eliminated, they can be managed through careful program design.

Physician "gaming" is a likely secondary effect of P4P. One of the more dramatic secondary effects of P4P was reported in an analysis of first year performance under the U.K.'s P4P program (20). Exclusion of patients from performance calculations was permitted as a form of risk adjustment, as long as exclusion criteria were followed. Doran et al. (20) found that physicians who excluded higher proportions of patients received more P4P reward monies. And, because a relatively large amount of "exception reporting" was concentrated in a small number of practices, there is at least the suggestion that some physicians were "gaming" the exclusion process.

Doran et al. (20) also reported that physician performance was lower "…in practices with a high proportion of patients who were living in single-parent or low-income households" providing some credence to concerns that practices serving low-income or disadvantaged populations might struggle under P4P programs. In contrast, in an analysis of data from Scotland, Sutton and McLean (66) found that practices located in "deprived" areas had higher quality scores. Their analysis also suggested that larger practices, with more clinicians, performed better on clinical quality indicators, while practices that received more money from sources other than the NHS performed less well. They concluded that the incentive effect of the P4P program was weaker for these latter practices.

Possible secondary effects that could be viewed as positive include: greater investment in electronic medical records systems by physician practices; an expanded role for nurses in the management of patients; greater numbers of physicians specializing in primary care (assuming P4P increases primary care physician incomes); and better documentation of care delivered in physician practices.

Emerging research suggests that better documentation is a secondary effect of P4P initiatives. There is evidence that one early effect of P4P in the U.S. and the U.K. may be better physician documentation of the care they provide in areas targeted by P4P (1, 25, 26, 30, 65). This is understandable, as better documentation may be the least costly action that physicians can take to improve their scores on quality indicators. Improving documentation of care is desirable for a variety of reasons, but P4P sponsors may be disappointed if it is the only outcome, and their award monies did not buy any actual increase in quality. Paying for improved documentation, however, may be a relatively short-lived phenomenon if physicians quickly exhaust opportunities to increase payments through this strategy.

Roland also reported that practices in the U.K. increased the number of nurses and other staff they employed, concurrently with the implementation of P4P (30). Physicians can add clinical and other expertise to their practices relatively quickly if there is the potential for that action to increase practice income. There also has been an upward trend in physician practice investment in electronic medical records in the U.K, but this probably was underway prior to P4P, stimulated by government reporting requirements (30).

Findings

Do the financial benefits of P4P outweigh the costs?

Only one study to date addresses this issue, and it shows a positive rate of return for an HMO incentive program.

Recent P4P program evaluations have focused primarily on quality impacts. However, given the various alternatives available to policy-makers to improve quality, it is reasonable to ask if P4P programs deliver benefits that exceed their direct costs. The costs of P4P programs include, at a minimum, the value of payouts to physicians and the costs of program administration. (A more challenging standard would require P4P programs to deliver net benefits that exceeded the net benefits of alternative approaches to quality improvement.)

Curtin et al. (17) attempted to address this question using evidence from an incentive program implemented within an HMO. The HMO rewarded adherence to treatment protocols in three clinical areas, as well as efficiency and patient satisfaction. The rewards program was carried out in the context of other attempts by the HMO to improve quality in these areas. The evaluators focused only on diabetes care, comparing projected treatment costs (trending forward past costs) to actual costs to estimate the benefits from the program.

The authors found a positive rate of return for the initiative, even though the program required providing additional services to diabetics. The authors acknowledged that their study design was limited by the lack of a contemporaneous control group, the short time period over which trends were calculated and results measured and the presence of other quality improvement efforts. The demand on the part of payers for evidence concerning the net financial benefits of P4P seems likely to increase if more P4P evaluations are able to document quality improvements. At the present time, however, this single study clearly cannot provide a definitive assessment of the rate of return from P4P programs. One would expect the rate of return to vary with type of condition and characteristics of the program setting, as well as the amount of monies paid to physicians.

Implications for Policy-Makers

Program evaluations indicate that P4P, when combined with other quality initiatives, is associated with quality improvement; however, the role of P4P in contributing to those improvements often is unclear. Nevertheless, the evaluations do provide some specific guidance for Medicare and Medicaid policy-makers as they design and implement P4P programs.

Budgeting for P4P

A critical design issue is whether to use predetermined quality benchmarks as a basis for paying physicians in a P4P program. While relatively simple to implement and to explain to participating physicians, paying specified amounts to all physicians or physician organizations that achieve quality benchmarks can result in a relatively unpredictable funding commitment and could lead to expenditures in excess of budgeted amounts. This is especially the case if accurate, timely data on physician performance are not available at the time the benchmarks and rewards are established.

Defining expectations regarding initial performance

Policy-makers should be aware that initial payments to physicians in a P4P program could yield relatively little actual improvement in quality, depending on the structure of the reward system adopted. The evaluations suggest that this can occur for two reasons. Where payments are made based on benchmarks, dollars could flow primarily to physicians who performed at the benchmark level of quality prior to program implementation. In effect, these payments represent a reward for past performance, and will not necessarily result in substantial quality improvement. To raise the overall level of quality in the initial years of public P4P programs, policy-makers may wish to consider rewarding improvement on quality measures, even though this raises the difficult question of whether it is appropriate to reward physicians whose performance on quality measures may still (after improvement) be low. Second, irrespective of whether measures of improvement or achievement are used in calculating rewards, observed improvement may reflect better physician documentation of care and not actual improvement in quality, at least in initial program years.

Allocating resources for program management

While there is likely to be pressure in public programs to spend a relatively large proportion of P4P funds on direct payments to physicians, there is evidence from existing programs that, with respect to raising the level of quality, "the devil is in the details." Specifically, adequate funds will need to be allocated initially for communication with physicians regarding how performance is measured and rewards are structured. Also, if insufficient funds are allocated to program administration, resulting in payment delays or inaccurate payments, the credibility of the program could suffer, potentially affecting physicians' willingness to invest in achieving quality goals.

Committing to ongoing surveillance

Some type of "gaming" of the P4P rules seems not only possible, but likely, under any set of rules governing P4P programs. Policy-makers will need to allocate administrative funds and effort to oversight and be prepared to take corrective actions where necessary to protect program legitimacy.

Implications for Policy-Makers

Establishing P4P Demonstrations

Because of the limited research evidence regarding P4P effectiveness under different program designs and reward structures, Medicare administrators should consider establishing and evaluating demonstration programs designed to systematically vary elements of program design that seem most likely to influence the nature and size of the impacts of P4P on physician behavior (21, 28). These arguably would include size of rewards, type and number of measures used, characteristics of the quality improvement strategies within which P4P is implemented and proportion of physician revenues affected. With respect to the latter, Medicare and Medicaid might consider partnering with private health plans in mounting demonstration programs. All demonstration programs, whatever their specific design features, should take place over a time period sufficiently long for physicians to invest in infrastructure changes needed to improve quality and for P4P program administrators to observe possible unintended impacts.

The Need for Additional Information

The literature on the response of physicians to financial rewards for providing better quality care is understandably sparse, although there is substantial interest in this topic among policy-makers. As a result, future research findings will be immediately useful in policy development. We believe that a research agenda should begin by addressing the following areas, with the understanding that new questions are likely to arise as the evidence base expands:

- Systematic documentation of the prevalence and characteristics of P4P-type programs for physicians, as well as their evolution over time.
- Expansion in the number and types of performance measures addressed by evaluations, in order to provide a more complete perspective on P4P effectiveness. Past evaluations have often been limited to a subset of measures used in P4P programs.
- Analysis of the factors that explain variation in physician responses (e.g., size and type of reward, practice characteristics, patient population characteristics) using different methodological approaches in order to enhance understanding of how, and under what circumstances, P4P is likely to be most effective.
- Analysis of the differential impact of financial incentives when they are combined with other quality improvement efforts.
- Tracking of physician responses to P4P programs over time, as performance measures and rewards change, to determine if short-term and long-term effects differ.
- Estimation of the net benefits of P4P efforts versus other possible approaches to improving quality, under different circumstances and for different patient conditions.

Appendix I References

1. Achat H, McIntyre P, Burgress M. "Health Care Incentives in Immunization." *Australian and New Zealand Journal of Public Health,* vol. 23, no. 3, 1999.

2. Amundson G, Solberg LI, Reed M, Martini EM, Carlson R. "Paying for Quality Improvement: Compliance with Tobacco Cessation Guidelines." *Joint Commission Journal on Quality & Safety,* vol. 29, no. 2, 2003.

3. Anonymous. "Healthcare Coverage: Survey Shows Pay for Performance Programs Continue to Increase." *Life Science Weekly,* 2005.

4. Armour BS, Pitts MM, Maclean R, Cangialose C, Kishel M, Imai H, Etchason J. "The Effect of Explicit Financial Incentives on Physician Behavior." *Archives of Internal Medicine,* vol. 161, no. 10, 2001.

5. Beaulieu ND, Horrigan D. "Putting Smart Money to Work for Quality Improvement." *Medical Care Research and Review,* vol. 40, no. 5, Part I, 2005.

6. Blumenthal D. "Employer-Sponsored Health Insurance in the United States—Origins and Implications." *The New England Journal of Medicine,* vol. 355, no. 1, 2006.

7. Blumenthal D. "Employer-Sponsored Insurance—Riding the Health Care Tiger." *The New England Journal of Medicine,* vol. 355, no. 2, 2006.

8. Bodenheimer T. "The American Health Care System. The Movement For Improved Quality In Health Care." *The New England Journal of Medicine,* vol. 340, no. 6, 1999.

9. Bodenheimer T, May JH, Berenson RA, Coughlan J. *Can Money Buy Quality? Physician Response to Pay for Performance.* Center for Studying Health System Change Issue Brief No. 102, 2005.

10. Bokhour BG, Burgess Jr JF, Hook JM, White B, Berlowitz D, Guldin MR, Meterko M, Young GJ. "Incentive Implementation in Physician Practices: A Qualitative Study of Practice Executive Perspectives on Pay for Performance." *Medical Care Research and Review,* vol. 63, no. 1, 2006.

11. Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B, Roland M. "Quality of Primary Care in England with the Introduction of Pay for Performance." *The New England Journal of Medicine,* vol. 357, no. 2, 2007.

12. Casalino LP, Alexander GC, Jin L, Konetzka T. "General Internists' Views on Pay-For-Performance and Public Reporting of Quality Scores: A National Survey." *Health Affairs,* vol. 26, no. 2, 2007.

13. Casalino L, Elster A. "Will Pay-for-Performance and Quality Reporting Affect Disparities?" *Health Affairs Web Exclusives,* vol. 26, no. 3, 2007.

14. Christianson J, Trude S. "Managing Costs, Managing Benefits: Employer Decisions in Local Health Care Markets." *Health Services Research,* vol. 38, no. 1, 2003.

15. Chung RS, Chernicoff HO, Nakao KA, Nickel RC, Legorreta AP. "A Quality-Driven Physician Compensation Model: Four-Year Follow-Up Study." *Journal for Healthcare Quality,* vol. 25, no. 6, 2003.

16. Conrad DA, Christianson JB. "Penetrating the "Black Box": Financial Incentives for Enhancing the Quality of Physician Services." *Medical Care Research and Review,* vol. 61, no. 3 Suppl, 2004.

17. Curtin K, Beckman H, Pankow G, Milillo Y, Greene R. "Return on Investment in Pay for Performance: A Diabetes Case Study." *Journal of Healthcare Management,* vol. 51, no. 6, 2006.

18. Damberg CL, Raube K, Williams T, Shortell SM. "Paying for Performance: Implementing a Statewide Project in California." *Quality Management in Health Care,* vol. 14, no. 2, 2005.

19. de Brantes F, Camillus JA. *Evidence-Based Care Rates: A New Health Care Payment Model.* Commonwealth Fund Publication No. 1022. New York, NY: The Commonwealth Fund, 2007.

20. Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M. "Pay-for-Performance Programs in Family Practices in the United Kingdom." *The New England Journal of Medicine,* vol. 355, no. 4, 2006.

21. Dudley RA. "Pay-for-Performance Research: How to Learn What Clinicians and Policy Makers Need to Know." *Journal of the American Medical Association,* vol. 294, no. 14, 2005.

Appendix I References

22. Dudley RA, Frolich A, Robinowitz DL, Talavera JA, Broadhead P, Luft HS, McDonald K. *Strategies to Support Quality-based Purchasing: A Review of the Evidence.* Technical Review Number 10, AHRQ Publication No. 04-0057, 2004.

23. Dudley RA, Rosenthal MB. *Pay for Performance: A Decision Guide for Purchasers.* Prepared for the Agency of Healthcare Research and Quality, Contract No. 290-10-0001/298. AHRQ Publication No. 06-0047, 2006.

24. Epstein AM, Lee TH, Hamel MB. "Paying Physicians For High-Quality Care." *The New England Journal of Medicine,* vol. 350, no. 4, 2004.

25. Fairbrother G, Hanson KL, Friedman S, Butts GC. "The Impact of Physician Bonuses, Enhanced Fees, and Feedback on Childhood Immunization Coverage Rates." *American Journal of Public Health,* vol. 89, no. 2, 1999.

26. Fairbrother G, Siegel MJ, Friedman S, Kory PD, Butts GC. "Impact of Financial Incentives on Documented Immunization Rates in the Inner City: Results of a Randomised Controlled Trial." *Ambulatory Pediatrics,* vol. 1, no. 4, 2001.

27. Felt-Lisk S, Gimm G, Peterson S. "Making Pay-for-Performance Work in Medicaid." *Health Affairs Web Exclusive,* vol. 26, no. 4, 2007.

28. Fisher ES. "Paying for Performance—Risks and Recommendations." *The New England Journal of Medicine,* vol. 355, no. 18, 2006.

29. Frolich A, Talavera JA, Broadhead P, Dudley RA. "A Behavior Model of Clinician Responses to Incentives to Improve Quality." *Health Policy,* vol. 80, no. 1, 2007.

30. Galvin R. "Pay-for-Performance: Too Much of a Good Thing? A Conversation with Martin Roland." *Health Affairs Web Exclusives,* vol. 25, no. 5, 2006.

31. Galvin R, Milstein A. "Larger Employers' New Strategies in Health Care." *The New England Journal of Medicine,* vol. 347, no. 12, 2002.

32. Grady KE, Lemkau JP, Lee NR, Caddell C. "Enhancing Mammography Referral in Primary Care." *Preventive Medicine* vol. 26, no. 6, 1997.

33. Greene RA, Beckman H, Chamberlain J, Partridge G, Miller M, Burden D, Kerr J. "Increasing Adherence to a Community-Based Guideline for Acute Sinusitis through Education, Physician Profiling and Financial Incentives." *The American Journal of Managed Care,* vol. 10, no. 10, 2004.

34. Grol R. "Improving the Quality of Medical Care: Building Bridges among Professional Pride, Payer Profit, and Patient Satisfaction." *Journal of the American Medical Association,* vol. 286, no. 20, 2001.

35. Grol R, Grimshaw J. "From Best Evidence to Best Practice: Effective Implementation of Change in Patients' Care." *Lancet,* vol. 362, no. 9391, 2003.

36. Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J, Lusk E. "Physician Financial Incentives and Feedback: Failure to Increase Cancer Screening in Medicaid Managed Care." *American Journal of Public Health,* vol. 88, no. 11, 1998.

37. 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,* vol. 104, no. 4 Part 1, 1999.

38. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. "The Unreliability of Individual Physician "Report Cards" for Assessing the Costs and Quality of Care of a Chronic Disease." *Journal of the American Medical Association,* vol. 281, no. 22, 1999.

39. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: National Academy Press, 2001.

40. Institute of Medicine. *Rewarding Provider Performance: Aligning Incentives in Medicare.* Washington, DC: National Academy Press, 2007.

41. Japsen B. "Medicare to Pay Doctors for Quality Information." *Chicago Tribune,* 2006.

Appendix I References

42. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH, LaForce FM. "Performance-Based Physician Reimbursement and Influenza Immunization Rates in the Elderly. The Primary-Care Physicians of Monroe County." *American Journal of Preventive Medicine,* vol. 14, no. 2, 1998.

43. Kuhmerker K, Hartman T. *Pay-for-Performance in State Medicaid Programs. A Survey of State Medicaid Directors and Programs.* Commonwealth Fund Publication #1018, New York, NY: The Commonwealth Fund, 2007.

44. Larsen D, Cannon W, Towner S. "Longitudinal Assessment of a Diabetes Care Management System in an Integrated Health Network." *Journal of Managed Care Pharmacy,* vol. 9, no. 6, 2003.

45. Levin-Scherz J, DeVita N, Timble J. "Impact of Pay-for-Performance Contracts and Network Registry on Diabetes and Asthma HEDIS Measures in an Integrated Delivery Network." *Medical Care Research and Review,* vol. 63, no. 1, 2006.

46. McGlynn EA, Asch SM, Adams J, Keesey J, Hicks J, DeCristofaro A, Kerr EA. "The Quality of Health Care Delivered to Adults in the United States." *New England Journal of Medicine,* vol. 348, no. 26, 2003.

47. McMahon LF, Hofer TP, Hayward RA. "Physician-Level P4P—DOA? Can Quality-Based Payment be Resuscitated?" *American Journal of Managed Care,* vol. 13, no. 5, 2007.

48. Morrow RW, Gooding AD, Clark C. "Improving Physicians' Preventive Health Care Behaviour through Peer Review and Financial Incentives." *Archives of Family Medicine,* vol. 4, no. 2, 1995.

49. Palsbo SE, Miller VP, Pan L, Bergsten C, Hodges DN, Barnes C. *HMO Industry Profile 1993 Edition.* Washington, DC: Group Health Association of America, Inc., 1993.

50. Pawson R, Tilley N. *Realistic Evaluation.* Thousand Oaks, CA: Sage Publications, Inc., 1997.

51. Peterson LA, Woodard LD, Urech T, Daw C, Sookanan S. "Does Pay-for-Performance Improve the Quality of Health Care?" *Annals of Internal Medicine,* vol. 145, no. 4, 2006.

52. Pink GH, Brown AD, Studer ML, Reiter KL, Leatt P. "Pay-for-Performance in Publicly Financed Healthcare: Some International Experience and Considerations for Canada." *HealthcarePapers,* vol. 6, no. 4, 2006.

53. Reschovsky J, Hadley J. *Physician Financial Incentives: Use of Quality Incentives Inches Up, but Productivity Still Dominates.* Issue Brief No. 108. Washington, DC: Center for Studying Health System Change, 2007.

54. Robinson JC. "Theory and Practice in the Design of Physician Payment Incentives." *Milbank Quarterly,* vol. 79, no. 2, 2001.

55. Roland M. "Linking Physicians' Pay to the Quality of Care—A Major Experiment in the United Kingdom." *The New England Journal of Medicine,* vol. 351, no. 14, 2004.

56. Rosenthal MB. "P4P: Rumors of Its Demise May Be Exaggerated." *American Journal of Managed Care,* vol. 13, no. 5, 2007.

57. Rosenthal MB, Dudley RA. "Pay-for-performance. Will the Latest Payment Trend Improve Care?" *Journal of the American Medical Association,* vol. 297, no. 7, 2007.

58. Rosenthal MB, Fernandopulle R, Song HSR, Landon B. "Paying for Quality: Providers' Incentives for Quality Improvement." *Health Affairs,* vol. 23, no. 2, 2004.

59. Rosenthal MB, Frank RG. "What is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review,* vol. 63, no. 2, 2006.

60. Rosenthal MB, Frank RG, Li Z, Epstein AM. "Early Experience with Pay-for-Performance: From Concept to Practice." *Journal of the American Medical Association,* vol. 294, no. 14, 2005.

61. Rosenthal MB, Landon BE, Normand S-LT, Frank RG, Epstein AM. "Pay for Performance in Commercial HMOs." *The New England Journal of Medicine,* vol. 355, no. 18, 2006.

62. Rowe JW. "Pay-for-Performance and Accountability: Related Themes in Improving Health Care." *Annals of Internal Medicine,* vol. 145, no. 9, 2006.

63. Sage WM, Kalyan DN. "Horses or Unicorns: Can Paying for Performance Make Quality Competition Routine?" *Journal of Health Politics, Policy and Law,* vol. 31, no. 3, 2006.

Appendix I References

64. Sidawy AN. "Pay for Performance: The Process and Its Evolution." *Journal of Vascular Surgery,* vol. 44, no. 4, 2006.

65. Simpson CR, Hannaford PC, Lefevre K, Williams D. "Effect of the U.K. Incentive-Based Contract on the Management of Patients with Stroke in Primary Care." *Stroke,* vol. 37, no. 9, 2006.

66. Sutton M, McLean G. "Determinants of Primary Medical Care Quality Measured under the New U.K. Contract: Cross Sectional Study." *British Medical Journal,* vol. 32, no. 7538, 2006.

67. Teleki S, Damberg C, Pham C, Berry S. "Will Financial Incentives Stimulate Quality Improvement? Reactions from Frontline Physicians." *American Journal of Medical Quality,* vol. 21, no. 6, 2006.

68. Town R, Kane R, Johnson P, Butler M. "Economic Incentives and Physicians' Delivery of Preventive Care. A Systematic Review." *American Journal of Preventive Medicine,* vol. 28, no. 2, 2005.

69. Town R, Wholey DR, Kralewski J, Dowd D. "Assessing the Influence of Incentives on Physicians and Medical Groups." *Medical Care Research and Review,* vol. 61, no. 3 Suppl, 2004.

70. Trude S, Au M, Christianson JB. "Health Plan Pay-for-Performance." *The American Journal of Managed Care,* vol. 12, no. 9, 2006.

71. Webber A. "Pay for Performance: National and Local Accountability for Health Care." *Managed Care,* vol. 14, no. 12 Suppl, 2005.

72. Williams TR. "Practical Design and Implementation Considerations in Pay-for-Performance Programs." *The American Journal of Managed Care,* vol. 12, no. 2, 2006.

73. Young GJ, Conrad DA. "Practical Issues in the Design and Implementation of Pay-for-Quality Programs." *Journal of Healthcare Management,* vol. 52, no. 1, 2007.

74. Young GJ, Meterko M, White B, Bokhour BG, Stautter KM, Berlowitz D, Burgess, Jr JF. "Physician Attitudes toward Pay-for-Quality Programs: Perspectives from the Front Line." *Medical Care Research and Review,* vol. 64, no. 3, 2007.

Appendix II Healthcare Effectiveness Data and Information Set (HEDIS)

HEDIS is a collection of standardized performance measures that the National Committee for Quality Assurance (NCQA) updates annually. Development of HEDIS 1.0 by a group of HMOs and several large national employers began in 1989 and took two years to complete. By 1992, however, NCQA assumed responsibility for HEDIS.[1,2] The standardized HEDIS measures allow comparisons across health plans in categories such as effectiveness of care, access and availability, health plan stability and management, cost, patient utilization, and satisfaction with the care experience. Approximately 560 health plans are expected to submit data for construction of HEDIS measures in 2007. While some report voluntarily, others, such as BlueCross BlueShield of Montana, are required by state law to report HEDIS measures.[3] HEDIS originated with HMOs, but efforts are underway to encourage commercial PPOs to voluntarily submit their HEDIS and CAHPS data to NCQA. In 2006, over 80 PPOs submitted data, more than twice the number of PPOs that submitted in 2005.

Effectiveness of care	Access and availability	Patient utilization	Satisfaction with the care experience	Plan stability and management	Cost of care
• Cancer screening • Cholesterol management for patients with cardio-vascular conditions	• Adult access to preventative/ ambulatory health services • Children and adolescents' access to primary care practitioners • Call answer timeliness and abandonment	• Impatient utilization—acute and nonacute care • Ambulatory care • Mental health utilization	• CAHPS 4.0H adult survey • CAHPS 3.0H child survey • Children with chronic conditions	• Board certification of physicians • State enrollment per product line • Diversity of membership	• Relative resource use for people with diabetes • Relative resource use for people with acute back pain • Relative resource use for people with COPD

1 Schoenbaum S. "What's Ahead on Quality: The Managed Care Perspective." *Physician Executive,* Nov-Dec, 1993.

2 National Chronic Care Consortium. State of the Art in Network Performance Measures. Available at: *http://www.nccconline.org/products/N32097.pdf.*

3 BlueCross BlueShield of Montana. Available at: *http://www.bcbsmt.com/Providers/providers_pub-hedis.asp0.*

Appendix III Descriptions of Selected Pay-for-Performance Programs

PacifiCare

PacifiCare Health Systems, a subsidiary of United Healthcare, has over three million plan members and nearly nine million specialty plan members.[1] PacifiCare contracts with about 300 large multispecialty physician organizations and has measured the performance of these groups on clinical and patient satisfaction indicators since 1993.[2] In 2002, the health plan implemented a new quality incentive program (QIP) for its California network, and in 2003 it began paying its physician organizations in California bonuses for meeting or exceeding ten clinical and patient satisfaction measures.[2] PacifiCare is a member of the Integrated Healthcare Association (IHA) and a participant in the IHA statewide pay-for-performance program.

Clinical	Patient experience	Information technology
• Diabetes management → Percent of members 18–75 with diabetes who had each of the following — HbA1c — HbA1c poor control >9.0% — LDL-C screening <100 mg/dl — Nephropathy monitoring • Breast Cancer Screening → Percent of women 40–69 with annual mammogram • Percent of adults 50–80 who had appropriate screening for colorectal cancer	• Doctor-patient communication — Can patient easily understand clinical guidance? — Did physician show respect for patient? • Timely access to care — Can patients get an appointment when they want one? — Do the appointments start on time? • Specialty Care — Can patients get an appointment to see a specialist if needed?	• Integrate clinical electronic datasets for population management — Computerized registries updated twice per year — Actionable reports/query lists from a disease registry updated twice per year • Support clinical decision making at point of care: — E-prescribing — E-drug checks for safety and efficiency — E-lab results — E-Reminders

1 PacifiCare company profile: *http://www.pacificare.com/commonPortal/link?navnode=CompanyProfiles.6.*

2 Rosenthal MB., Frank RG, Li Z, Epstein AM. "Early Experience with Pay-for-Performance: From Concept to Practice." *Journal of the American Medical Association,* vol. 294, no. 14, 2005.

Appendix III Descriptions of Selected Pay-for-Performance Programs

BlueCross BlueShield of Michigan

BlueCross BlueShield of Michigan (BCBSM) began operating the Physician Group Incentive Program (PGIP) in January 2005. The program targeted three components of the delivery process for improvement: chronic illness treatment, prescribing patterns for BCBSM members, and physician participation in care management and shared decision-making programs.[1]

In quarter two of 2006, BCBSM launched its Physician Organization Gain Sharing program (POGS) to "strengthen the performance improvement infrastructure available to clinicians" and to achieve measurable savings in the following care delivery areas: pharmacy, laboratory, diagnostic imaging, in-network referrals and hospitalizations.[2,3] In the POGS, physician organizations receive defined payments based on the size of their BCBSM membership, and BCBSM commits to sharing at least 50 percent of overall program savings with selected physicians.[3]

As of January 2007, the PGIP and POGS Incentive Programs (combined) include 31 physician groups comprised of 5,500 physicians treating 1,277,000 patients. BCBSM paid out $19.9 million in incentives in CY2006 and reported a generic dispensing rate increase of 1.5 percent from CY 2005 to 3rd quarter 2006 and forecast a four percent increase for 2007.[3]

1 Collaborating for Quality Improvement: BCBSM Value Report. October 2005. Available at: *http://www.bcbsm.com/pdf/collaborating_quality_improvement.pdf.*

2 Phone call with Mark Casmer, Manager for Clinical Program Development at BCBSM.

3 Value Partnership: Physician Incentive Programs Update. BCBSM Liaison Group. March 2007.

Highmark BlueCross BlueShield

Highmark, which serves 29 counties in western Pennsylvania, initiated its Quality Blue P4P program in 1996 for primary care physicians, focusing on reductions in the total cost of care. As managed care continued to decline, the health plan began reducing the emphasis on "cost of care," replacing it with quality measures. By 2005, Highmark's revised P4P methodology was utilized for both for FFS and capitated members.[1] There are a total of 14 HEDIS-like quality measures. The performance of physicians participating in the Quality Blue program is evaluated relative to the performance of their peers practicing in the same specialty. Participants are awarded one point if they meet or exceed the specialty average for a clinical category and 0.5 points for meeting or exceeding 90 percent of the specialty average. The program also awards points for higher generic prescribing rates, expanded office hours, use of EHR and e-Rx technology and quality improvement projects or certifications like the NCQA and ABIM PIM's.

Bonus payments are made as add-ons to evaluation and management services billed by the primary care physician. Physicians receive payments of $0 to $9 based on their performance on a real time basis at the time the claim is processed.[2]

As of April 2007, Highmark's physician incentive plan included only primary care providers. Highmark expects to grow the program to include ten clinical specialties by 2009.[2] This initiative will begin in 2007 with profiling only, no P4P. Public reporting of the specialist data will begin in 2008. This phase-in period will provide physicians the opportunity to improve group performance and permit physician feedback to Highmark prior to the implementation of the rewards program.[3]

1 The Highmark Story. Available at: *https://www.highmark.com/hmk2/about/index/shtml.*

2 Quality Blue: A Physician Pay-for-Performance Program. Highmark BCBS.

3 Phone call with Dr. Michael Madden, MD, Medical Director for Quality and Medical Performance Management at Highmark BCBS.

Appendix III Descriptions of Selected Pay-for-Performance Programs

Bridges to Excellence

Bridges to Excellence, or BTE, is a nonprofit organization that administers monetary rewards and recognition programs for providers: Diabetes Care Link, Cardiac Care Link (CCL) and Physician Practice Connection. These programs are based on NCQA clinical performance measures. Physicians contracting with participating health plans, employer coalitions or employers can receive bonuses of up to $160 per eligible cardiac patient under the CCL program, up to $80 per eligible diabetic patient and up to $50 per eligible member under the physician office link reward program, although the potential reward varies depending on the physician's local incentive program payment methodology.[1]

Beginning in 2007, BTE will implement the Spine Care Link (SCL) and the Internal Medicine Care Link (IMCL). Under the SCL program, participating physicians could be eligible for bonuses of up to $50 per member per year and receive BTE certification based on relative performance level.[2] The IMCL program, a partnership with the American Board of Internal Medicine (ABIM), will allow internists to qualify for maintenance of board certification, continuing education credits and bonus payments. As of April 2007, these scoring and payment methodologies were not yet estabilished.[3]

Diabetes care link	Cardiac care link	Physician office link
• HbA1c control — <7.0% → 40% of patients; — >9% → ≤15% of patients • LDL control — ≤130 mg/dL → ≤37% of patient — <100mg/dL → 36% of patients • Blood pressure control — ≥140/90mmHg → ≤35% patients — <130/80mmHg → 25% patients • Eye examination → 60% patients • Nephropathy assessment → 80% of patients • Foot Examination → 80% of patients • Smoking status and cessation advice or treatment → 80% of patients	• Blood pressure control <140/90 mmHg → 75% of patients • Complete lipid profile → 80% of patients • Cholesterol control <100 mg/gL → 50% of patients • Administering aspirin or other antithrombotic → 80% of patients • Smoking cessation counseling or treatment	• Access and communication • Patient tracking and registry • Care management • Patient self-management support • Electronic prescribing • Test tracking • Referral tracking • Performance reporting and improvement • Interoperability

1 AHA/ASA/NCAQA Heart/Stroke Recognition Program: Standards, Performance Criteria, and Scoring. Available at: *http://web.ncqa.org/tabid/140/Default.aspx*.

2 *http://www.bte.org.*

3 Email correspondence with Sarah Burstein, Project Leader, Bridges to Excellence.

Appendix III Descriptions of Selected Pay-for-Performance Programs

Integrated Healthcare Association

The Integrated Health Association (IHA), a collaborative nonprofit organization based in California, initiated a pay-for-performance program in 2002.[1] As of 2006, seven health plans representing over 6 million commercial HMO/PPO members participated in the program.[2] The performance measures fall into four categories: clinical, patient experience, IT enabled systems and efficiency, all of which are adapted from the NCQA's Healthcare Effectiveness Data and Information Set (HEDIS).[3] Each year, between July and October, payments commensurate with physician group performance are distributed to each group by the participating health plans.[4] Actual payment versus the total potential payout varies for each health plan but is a function of the relative performance of the group. A ten percent physician bonus opportunity is available if the physician organization operates an internal physician monitoring system to measure patient satisfaction and has evidence-based clinical quality measures, and also if other incentives of monetary value are in place, such as for attendance at professional conferences.

Clinical	Patient experience	Information technology	Physician incentive bonus
• Diabetes management → Percent of members 18-75 with diabetes who had each of the following — HbA1c — HbA1c poor control >9.0% — LDL-C screening <100 mg/dl — Nephropathy monitoring • Breast Cancer Screening → Percent of women 40-69 with annual mammogram • Percent of adults 50-80 who had appropriate screening for colorectal cancer	• Doctor-patient communication — Can patient easily understand clinical guidance? — Did physician show respect for patient? • Timely access to care — Can patients get an appointment when they want one? — Do the appointments start on time? • Specialty Care — Can patients get an appointment to see a specialist if needed?	• Integrate clinical electronic datasets for population management — Computerized registries updated twice per year — Actionable reports/query lists from a disease registry updated twice per year • Support clinical decision making at point of care: — E-prescribing — E-drug checks for safety and efficiency — E-lab results — E-reminders	• Physician organizations may qualify for physician incentive bonus opportunity if incentives are provided to individual physician for: — Clinical quality — Patient experience • Individualized rewards may include — Bonus paid commensurate with performance — Tangible rewards to physicians, such as increased patient volume

1 Damberg, CL, Raube K, Williams T, Shortell SM. "Paying for Performance: Implementing a Statewide Project in California." *Quality Management in Health Care,* vol. 14, 2005.

2 *http://www.iha.org.*

3 Integrated Healthcare Association California Pay-for-Performance Program: P4P Measurement Year 2006 Manual. Updated November 30, 2006.

4 Email correspondence with Delores Yanagihara, IHA P4P Program Development Manager.

Appendix III Descriptions of Selected Pay-for-Performance Programs

United Kingdom

Over a period of 18 months, the British Medical Association and the National Health Service Confederation, with the assistance of a small group of academic advisors, negotiated a pay-for-performance program for primary care physicians. These practitioners can earn up to 1,000 points for performance on a set of indicators, with an additional 50 points for provision of prompt access to services. The indicators relate to clinical care in ten areas (550 points), practice organization and patient experience. The formula for assigning points takes into account practice characteristics, and criteria are specified for excluding individual patients from the calculations.[1,2] Approximately $3.2 billion in new funds was allocated for physician rewards to be distributed over a 3-year period. Payment was limited to $133 per point in 2004–2005, with a maximum possible payment of $139,400 per physician. In its first year, the median practice achieved 95.5% of available points. On average, gross practice income grew by about $40,000 for that year.[3,4]

Clinical	Organizational
• Coronary heart disease • Stroke, Transient ischemic attack • Hypertension • Hyperthyroidism • Diabetes • Mental disorder • Chronic obstructive pulmonary disease • Asthma • Epilepsy • Cancer	• Records and information about patients (e.g., smoking status recorded) • Communicating with patients (e.g., availability of staff to talk with patients by phone) • Education and training (e.g., documented minimum number of reviews of significant events) • Management of medications (e.g., medication review of patients within a specific time period) • Management of practice (e.g., backing up computer data)

1 Roland M. "Linking Physicians' Pay to the Quality of Care—a Major Experiment in the United Kingdom." *The New England Journal of Medicine,* vol. 351, no. 14, 2004.

2 McDonald R, Harrison S, Checkland K, Campbell SM, Roland M. "Impact of Financial Incentives on Clinical Autonomy and Internal Motivation in Primary Care: Ethnographic Study." *British Medical Journal,* vol. 334, 2007.

3 Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald,B, Roland M. "Quality of Primary Care in England with the Introduction of Pay-for-Performance." *The New England Journal of Medicine,* vol. 357, 2007.

4 Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M. "Pay-for-Performance Programs in Family Practices in the United Kingdom." *The New England Journal of Medicine,* vol. 355, no. 4, 2006.

Appendix IV Methodological Discussion

The studies reviewed or cited in this Synthesis use a variety of methodological approaches. The purpose of this Appendix is to provide a non-technical overview of these approaches, addressing which are likely to be the most useful and appropriate for addressing different types of research questions.

Experimental Approaches

When reviewing clinical literature and, increasingly, health services research literature, authors frequently order studies according to their "strength of evidence," where the strongest, most reliable findings typically are presumed to be generated by studies employing some type of experimental design. In many cases, in fact, studies that do not employ experimental designs are excluded from reviews because their findings are not considered, a priori, to be reliable. The implication is that non-experimental studies represent "less rigorous" research and, as such, do little to advance the field of research, and that findings from non-experimental studies should be heavily discounted or ignored by policy-makers. This is unfortunate because, while experimental studies are certainly preferable for answering specific types of questions, they are less useful in addressing others. For instance, in evaluating the impact of financial incentives on the quality of care delivered by physicians, experimental designs are limited in their ability to address many questions of great interest to policy-makers, such as: Why do physician incentives change physician behavior in some circumstances and not others? How should incentives be structured to achieve their greatest impact? To understand the strengths and limitations of the evidence generated by the experimental studies cited in this synthesis, it is important to be clear about what experimental methods are intended to accomplish.

Basically, experimental designs are intended to eliminate all factors that could affect an outcome except for the "intervention" of interest. For a pay-for-performance program or similar initiative that uses financial incentives in an effort to improve the quality of care provided by physicians, the most effective way to do this, theoretically, would be to randomly assign physicians into two groups, with one group "exposed" to the incentive and a second group not exposed. The outcome(s) of interest to policy-makers, and that is expected to change as a result of the intervention, then is measured for each group before and after the intervention. Post intervention measurement is carried out after enough time has elapsed to allow the intervention to have an impact, with the change in outcome measure calculated for each group. If a difference in these two "change measures" is observed that is so large it was unlikely to have occurred by chance, the analyst then concludes it resulted from the intervention—in this case, the financial incentive to which the first group of physicians was exposed. An important aspect of experimental designs is that they are not constructed to shed light on the "mechanism" by which the change occurred. That is, what physician behaviors changed to generate the findings? How did they change? Why were changes observed for some physicians and not others? For some outcome measures and not others? In fact, the goal of the experimental method is to eliminate any possible differences between the two groups of physicians, and it is these differences that are likely to be essential in understanding the change process.

Obviously, there are severe practical obstacles to employing "pure" randomized designs in assessing the impact of physician financial incentives on quality of care. Dudley (21) discusses several of these obstacles. Randomization processes are hard to carry out rigorously in real world settings and, in part to assure rigorous randomization, the settings chosen may be limited in their comparability to most physician practices. To control costs and achieve cooperation on the part of

physicians, the "sample" of physicians or physician practices is often quite small, making it hard to determine if any observed differences in outcomes are due to the incentive program, or simply occurred by chance.

The authors of the experimental studies cited in this synthesis were well aware of these difficulties and limitations. When they found no significant impacts–which was the case most of the time, as we note in the Synthesis–the authors typically offered relatively ad hoc, but nevertheless very persuasive, explanations of why that was the case. Or, they retrospectively attempted to collect information on physician behaviors that may have worked against finding an impact from the incentive. Even when significant differences were detected, some authors sought to understand this result through subsequent data analysis outside of the experimental design and, in one case, concluded that it was due to better documentation rather than actual improvements in care (25, 26).

In an attempt to avoid some of the limitations of pure randomized designs, other authors evaluated the impact of ongoing incentive programs implemented by health plans or other payers. Randomization of physicians did not occur in these programs, but researchers sought to retain the strengths of an "experimental" approach by constructing "quasi-experimental" designs. The challenge to the researchers is to identify a "comparison group" that has characteristics, which mirror, as closely as possible, the "non-intervention" group in a randomized design. This often requires some ingenuity. Rosenthal et al. (60) used this approach, designating physician groups in a different region served by a health plan, and where an incentive was not in effect, as the "non-intervention" group in their study.

No comparison group in a quasi-experimental design is ever ideal, however, in the sense of perfectly replicating what would have occurred through a randomization process. To address this problem, researchers attempt to gather as much data as possible on different characteristics of intervention and comparison group members, and then use statistical techniques to "adjust away" differences between the groups. Even then, there is the possibility that the groups differ on characteristics that are either not observable or that, while observable in theory, cannot be measured with existing data sources. As a consequence, when interpreting their results, researchers using quasi-experimental designs typically are careful to note that any differences they observe could reflect underlying differences in the groups and not necessarily the impact of the "intervention" being assessed. Thus, Rosenthal et al. (60) noted that they "…relied on the assumption that absent the [program], trends (or differences in trends) in quality improvement in California would have resembled those in the Pacific Northwest network. Although this assumption is generally supported by the similarity of [pre-program] trends between the two networks, it is not directly testable". Similarly, when no significant outcomes are found, there is a possibility that real differences between the groups may, in fact, exist but were obscured by unmeasured, uncontrolled differences in group characteristics. Nevertheless, quasi-experimental studies addressing actual physician incentive programs in real world settings generate findings that likely have greater "face validity" for policy-makers than findings from small scale, more tightly structured studies based on pure randomized designs. It is important to reiterate that, because quasi-experimental studies attempt to replicate, to the greatest degree possible, the randomized, experimental approach, they also are not designed to address the mechanisms by which changes in outcomes occurred, or the factors that inhibited or facilitated change. In discussing and interpreting their findings, researchers using quasi-experimental designs face the same challenges as those using randomized designs.

Appendix IV Methodological Discussion

The "before-after" approach is relatively common in evaluation literature of all types and used in some studies cited in the Synthesis. It typically is viewed as weaker than quasi-experimental designs in its ability to detect the impact of an intervention, like a physician incentive program, on an outcome measure of interest. In before/after studies, no external comparison group is used to track performance on the outcome measure before and after the intervention. Instead, the outcome of interest is tracked only for the intervention group, before and after the intervention. Essentially, the pre-intervention outcomes are taken to represent what would have occurred in the absence of the intervention. For example, the performance of physicians on a set of quality indicators, measured before the physicians are exposed to new financial incentives, and trended forward, might be used to represent the outcomes that would have occurred in the absence of the incentive program. Sometimes this is referred to as "using the participants as their own controls." The credibility of this approach is strongest in relatively stable environments, and where there are a relatively large number of pre-intervention observations of an outcome measure over time, so that statistical methods can reliably fit a time trend to these data.

If there are other factors in the environment that influence the outcome measures of interest, then the adoption of past performance as a reflection of what outcomes would have been in the absence of the intervention is less defensible. Concerns about this are certainly relevant when using a "before-after" approach to assess the impact of physician incentives on quality. There are a large number of ongoing efforts to improve quality, including organizational quality improvement efforts and public reporting of comparative quality data, and these efforts change over time. Statistical methods incorporating time trends can be used to control, in a relatively crude way, for the presence of these other efforts, but this approach is not always convincing. Essentially, the use of "before-after" analysis in assessing the impact of physician incentive programs to improve quality is less desirable than analysis that incorporates contemporaneous comparison groups in a quasi-experimental design. Researchers adopting the before-after approach recognize its limitations and often attempt to address them by comparing their findings to national trend data. Alternatively, they may restrict their analysis to relatively short time periods (e.g., one observation period before and one after implementation of an incentive program) to minimize the potential impact of external environmental changes and the need to fit statistical trend lines to past performance data. However, immediate responses to incentives could be quite different than longer-term responses. Again, because the before-after design, carefully implemented, essentially attempts to replicate a pure experimental approach to the greatest degree possible within the study setting, it has the same inherent limitations as noted above with respect to the questions it addresses.

Non-Experimental Approaches

Non-experimental approaches employ observational data to explore factors that influence the nature of physician responses to financial incentives and how those responses are, in turn, linked to the quality measures of interest. The observational data can be gathered through interviews, surveys, or existing (typically administrative) sources. In the best of these studies, attention is focused on understanding the mechanism of change, which typically requires an implicit or explicit "behavioral model" of individual or organizational response to the intervention. (For a discussion of this point see Frolich et al. (29) and Town et al. (69).) This behavioral model is used by researchers to generate hypotheses concerning which factors are likely to affect behavior and how the intervention is likely to interact with these factors to produce change. A "treatment/control" group comparison is not required. Rather than trying to eliminate differences through

randomization, the testing of hypotheses generated through behavioral models relies on, and exploits, existing variation on the characteristics of participants and the settings in which they respond to interventions. (This distinction is cogently drawn out by Pawson and Tilley (50).) For example, a researcher might hypothesize that physicians in large medical groups will improve performance to a greater degree than physicians in smaller groups. The model underlying this hypothesis might incorporate assumptions about the importance of infrastructure support in minimizing response costs, the way in which response costs are incurred by physicians in groups with different features and the relative importance of response costs in influencing behavior. The behavioral model would be relied on to identify the variables of interest and would guide data collection. However, in order to test the hypothesis, and explore its underlying structure, variation in the key variables is necessary. Consequently, the researcher would design the study to ensure that, to the extent possible, this variation existed (50). This contrasts with experimental research approaches, where a key part of the design involves eliminating differences in intervention and control groups. Both approaches are appropriate; they simply are designed to address different questions.

One key to mounting a non-experimental study or a stream of research that generates credible results on questions that are of interest to policy-makers is the development of a persuasive theory. This theory must permit the specification of hypotheses that can be tested and rejected. A common criticism of non-experimental studies is that they too often are little more than statistical "fishing expeditions" uninformed by theory. By this, critics usually mean that the researcher has used statistical techniques to analyze large amounts of data, found statistically significant relationships, and engaged in a variety of "after the fact," ad hoc speculations about the possible reasons for these relationships. In the worst case, the researcher may not have adjusted the statistical testing procedure, ignoring the likelihood that, when a large number of tests are conducted, some statistically significant relationships will surface by chance.

Ideally, in non-experimental studies, the results from the testing of hypotheses in one study would be used to revise the behavioral model and inform the design of subsequent physician incentive programs, generating opportunities to test new, refined hypotheses. In this manner, the mechanisms by which financial incentives directed at improving quality affect physician behavior could be better understood (21, 50). As a result, policy-makers could improve the effectiveness of pay-for-performance and other financial incentive programs over time. Ongoing research on the United Kingdom's physician pay-for-performance program holds this potential. No pre-program data were available on a consistent basis (although Campbell, et. al., 2007, and Simpson, et. al., 2006 used pre-program data on selected practices in their evaluation (11, 65)) and the program was instituted across all regions of the National Health Service, precluding the use of experimental methods to analyze program impacts. However, a non-experimental analysis of physician responses in the first year suggests that the willingness and ability of physicians to exclude complicated patients from the measurement process was a significant factor in achieving higher performance scores (20). This finding focuses the attention of policy-makers on refining patient exclusion criteria going forward.

Just as there are challenges in mounting "pure" experimental studies, designing and conducting non-experimental studies that proceed in a logical, sequential manner is likely to be daunting in practice. Most non-experimental studies are "one-off," as are most experimental studies. (An exception in the Synthesis is offered by the experimental studies conducted by Fairbrother et al. (25, 26).) Researchers take advantage of a particular incentive program mounted by payers

to analyze a specific question. Results seldom are used to restructure programs, with follow-on evaluations. Also, at least to date, analyses of physician responses to financial incentives intended to improve quality have generally not been guided by strong behavioral models. As a result, the research literature has not evolved in a coherent path that maximizes its usefulness to policy-makers. Dudley (21) observes that, for this to occur, funders would need to support "sequential and parallel hypothesis testing" (p. 1823) and argues that this is essential if research is to help policy-makers "…to understand the nuances of when and how incentives work" (p. 1823).

Appendix V Summaries of Controlled Experiments

Controlled experiments in the U.S. providing financial rewards to physicians for quality

Author	Geographic scope	Physicians	Type of data analyzed	Quality measure(s)	Financial incentives	Effect of financial incentives	Comments
Fairbrother et al., 1999 (25)	New York City	60 inner-city, office-based pediatricians randomly assigned to 1 of 3 interventions: bonus and feedback, enhanced fee-for-service and feedback, and feedback only	Chart reviews of 50 randomly selected children (3–35 months olds) at 4 month intervals over 1 year	Percent of children with up-to-date immunizations	Bonus: $1,000 per physician for a 20% improvement from baseline; $2,000 for a 40% improvement, and $5,000 for reaching 80% compliance. Enhanced FFS: $5 per vaccine administered on schedule, $15 per visit at which more than one scheduled vaccine was administered	Significant improvement in bonus group but not in other groups	Improvement was due primarily to better documentation
Fairbrother et al., 2001 (26)	New York City	57 physicians completed second study year; 24 in a group receiving bonus payments, 12 in enhanced fee-for-service, and 21 in a control group receiving performance feedback only	Chart reviews of 50 randomly selected children (3–35 months old) at 4 month intervals over 1 year	Percent of children with up-to-date immunizations	Bonus: At each data collection point, $1,000 for 30 percentage point improvement and $2,500 for 45 percentage point improvement; $5,000 for reaching 80% coverage and $7,500 for 90%. FFS: $5 per vaccine administered on schedule and $15 for each visit in which all due vaccines were administered	Significant improvement under both programs	Improvement was due primarily to better documentation
Grady et al., 1997 (32)	Dayton, OH and Springfield, MA	61 practices, with 1–6 community-based primary care physicians per practice, randomly assigned to one of three interventions: small incentive in combination with feedback; education; and education plus chart reminder	Chart audits of women 50 years and older conducted at baseline and quarterly for one year	Mammography referral rates, mammography completion rates, mammography compliance rates	Annual payment based on level achieved; amount not clear but called “token” by authors	None	Financial incentive was combined with feedback program

Controlled experiments in the U.S. providing financial rewards to physicians for quality (cont'd)

Author	Geographic scope	Physicians	Type of data analyzed	Quality measure(s)	Financial incentives	Effect of financial incentives	Comments
Hillman et al., 1998 (36)	Philadelphia	52 largest primary care sites affiliated with Medicaid HMO were randomly assigned to intervention or usual care	Chart audits at baseline and every 6 months for 1.5 years for women 50 years and older	Percent of charts in compliance with cancer screening guidelines for four types of cancer	The 3 sites with highest scores received bonus equal to 20% of capitation for target patient group; 3 practices with next highest scores and 3 with greatest improvement received 10% bonus	None	Physician awareness of intervention was 67%; with 30% of physicians not responding to a survey; screening rates for both groups approximately doubled in study period but remained low overall
Hillman et al., 1999 (37)	Philadelphia	49 physician practices in a Medicaid plan randomly assigned to feedback and incentive (19), feedback only (15), and control (15)	Chart audits at baseline and 6 month intervals over 18 months	Compliance with pediatric preventive care guidelines	3 sites with highest compliance scores received a bonus of 20% of the site's total 6-month capitation payment for children to age 7. The next 3 highest scoring sites received 10% of capitation, as did 3 sites showing most improvement	None	Only 56% of the practices randomized to the incentive program were aware of the program, but there was no difference in performance between practices that were aware and those that were not
Kouides et al., 1998 (42)	Monroe County, NY	54 solo or group practices that had participated in 1990 Medicare Demonstration Project, randomized to a control group and an incentive group	Physician-reported immunizations for elderly patients combined with public health immunizations and union-sponsored immunizations	1990 influenza immunization rate and improvement in rate from 1990–1991	Payment of $.80 per immunization if rate of 70% achieved; $1.60 per immunization if rate of 85% achieved	No effect on mean immunization rate; significantly larger improvement in immunization rate for incentive group	Financial incentive was combined with intensive promotion of immunizations at physician practices

Appendix VI Summaries of Program Evaluations

Evaluations of programs in the U.S. that provide financial incentives to physicians for quality

Author	Geographic scope	Physicians	Type of data analyzed	Quality measure(s)	Financial incentives	Effect of financial incentives	Comments
Amundson et al., 2003 (2)	Minnesota	Physicians in 19–20 medical groups participating in a network model HMO	Audits of 14,489 ambulatory patient records from 1996–1998	Documentation of tobacco use and discussion of tobacco use, with medical group targets of 80% for each	Bonus pools established for each medical group, with a portion of bonus payment directed to performance on tobacco quality measures	Documentation increased significantly for 13 of 20 groups, and discussion improved for 7 groups	No contemporaneous comparison group was present and increases in "discussion" may be due in part to better documentation
Beaulieu and Horrigan, 2005 (5)	Upstate New York	21 physicians and 624 diabetic patients	Performance self-reported by physicians three times in study year	Composite measure of performance in delivering diabetes care according to best practices	Physicians receive $3 pmpm for Medicare patients and $.75 for commercial for composite score above 6.86; $1.50 and $.37 for a score above 6.23 and $.75 and $.18 for a 50% improvement, with score below 6.23	Composite scores increased by 48%	It is not possible to separate the effect of other changes introduced at the same time from the effect of financial incentives, and physician participants were volunteers
Chung et al., 2003 (15)	Hawaii	800 of approximately 1,500 eligible physicians in a PPO volunteered to participate in first year	Claims data from 1997–2000	Use of ACE inhibitors or ARB in heart failure, measurement of HbAlc in diabetes, and rates of childhood immunizations	3.5% of base fees were earned, on average, by participating physicians; 17 physicians received maximum rewards in 2001, with 14 receiving $10,000 and 3 receiving $13,000	Consistent, statistically significant improvement in use of ACE inhibitors or ARB and in measurement of HbAlc	No contemporaneous control group
Felt-Lisk et al., 2007 (27)	California	Physician practices in five different Medicaid managed care plans that received payments for performance and practices in two plans that did not	Encounter data from 2002–2005, Medicaid administrative data, meeting notes, and interviews	Percentage of plan members meeting HEDIS well-baby visit guidelines	From 2003–2005, four of five plans paid bonuses to contracting practices based on the proportion of children who met well-baby visit guidelines. The fifth made payments directly to physicians from a bonus pool	There was overall improvement in performance related to guidelines for well-baby care but there was no effect in two plans, possible small effects in two plans. There appeared to be substantial improvement in only one plan	Little information was provided regarding statistical methods used to estimate difference in differences effects. More successful programs offered greater rewards and had better communication with physicians about program characteristics
Greene et al., 2004 (33)	Rochester, NY	500 internists, 200 family practitioners, and 200 pediatricians	Medical claims organized using episode treatment group methods	Treatment exceptions per episode in treatment of sinusitis	Amount withheld from capitation payment decreased from 15% to 10% for top 5% of performers and increased to 20% for the bottom 5%	Mean overall exceptions per episode decreased due to decrease in use of less effective or inappropriate antibiotics	The financial incentives were part of a multi-faceted intervention and it is not possible to determine the specific effect of financial incentives; no contemporaneous control group

Appendix VI Summaries of Program Evaluations

Evaluations of programs in the U.S. that provide financial incentives to physicians for quality (cont'd)

Author	Geographic scope	Physicians	Type of data analyzed	Quality measure(s)	Financial incentives	Effect of financial incentives	Comments
Larson, Cannon and Towner, 2006 (44)	Utah	400 employed physicians who were part of integrated health system	Laboratory data, health plan claims, physician billing, clinical information systems	Six different performance measures related to the treatment of diabetes	Overall financial incentive totaled 0.5% to 1.0% of compensation with about half directed at diabetes care	There was statistically significant improvement in all six indicators	Financial incentives were part of a complex, multifaceted quality improvement intervention, so it is not possible to determine the effect of the financial incentives by themselves; no contemporaneous control group
Leven-Scherz, DeVita and Timble, 2006 (45)	Massachusetts	Physicians participating in a provider network that contracts with health plans are compared to physicians not in the network	Medical claims	Performance relative to benchmarks for diabetes and asthma care, selected from larger group of performance measures	Bonus payments and the return of portion of withholds in managed care contracts; magnitude of the incentives not stated	Significantly greater improvement in diabetes measures, relative to the comparison group; no significant improvement in asthma care	It is not possible to isolate the effect of financial incentives from other quality improvement efforts implemented at the same time
Morrow et al., 1995 (48)	Northeastern United States	Primary care physicians contracting with an IPA model HMO	Audited medical chart data	Rates of childhood MMR immunization, cholesterol screening for adults, appropriate charting of information	Payment in addition to base capitation payment to primary care physicians; amount of payment for quality indicators not specified	Significant improvements in all measures	Longitudinal study with no contemporaneous control group
Rosenthal et al., 2005 (60)	California and the Pacific Northwest	Analytic sample included 134 medical groups contracting with a health plan in California that were exposed to a financial incentive and 33 groups in the Northwest contracting with the same groups but not exposed to the incentive	Claims-based performance data aggregated to the physician group level	Cervical cancer screening, mammography, and hemoglobin testing (selected from 10 measures subject to financial incentives)	Payments of $0.23 PMPM for each target achieved. A group with 10,000 plan members could potentially earn $270,000 per year	Significant improvement in cervical cancer screening relative to the control group; no significant improvement in the other two measures	Most (75%) of the dollars were earned by groups that had achieved the benchmarks prior to the incentive program, however there was substantial improvement in low performance groups

Notes

Notes

Notes

For more information about the Synthesis Project, visit the Synthesis Project Web site at *www.policysynthesis.org*. For additional copies of Synthesis products, please go to the Project's Web site or send an e-mail message to *pubsrequest@rwjf.org.*

PROJECT CONTACTS

David C. Colby, Ph.D., the Robert Wood Johnson Foundation

Brian C. Quinn, Ph.D., the Robert Wood Johnson Foundation

Claudia H. Williams, Synthesis Project

Sarah Goodell, Synthesis Project

Robert Wood Johnson Foundation

THE SYNTHESIS PROJECT

NEW INSIGHTS FROM RESEARCH RESULTS

RESEARCH SYNTHESIS REPORT NO. 13
DECEMBER 2007

The Synthesis Project
The Robert Wood Johnson Foundation
Route 1 & College Road East
P.O. Box 2316
Princeton, NJ 08543-2316
E-Mail: synthesisproject@rwjf.org
Phone: 888-719-1909

www.policysynthesis.org