Inventory and Literature Review: Pay-for-Performance Methods and Structure

Prepared for Minnesota Department of Health by
The University of Minnesota, under contract with MN Community Measurement

There is a developing literature on pay-for-performance initiatives directed at health care providers in the United States and other countries (Christianson, Leatherman and Sutherland, 2008a,b; Christianson, Leatherman and Sutherland, 2007). Under these initiatives, providers are paid more if they achieve quality benchmarks or demonstrate improvements in the quality of care they provide. Although not common, some initiatives involve financial penalties for providers that fail to achieve quality targets or demonstrate improvement.

In pay-for-performance initiatives, quality is measured in a variety of ways, including: presence of certain characteristics in the practice environment (e.g. maintenance of a patient registry for people with specific chronic illnesses, or use of electronic medical records); carrying out desirable practice activities (e.g. initiating regular contact with patients who have chronic illnesses); conforming to evidence-based practice recommendations (e.g. for diabetics, meeting recommendations for frequency of blood sugar testing), or achieving goals for biologic measures (e.g., for diabetics, blood sugar levels below a specified target.) Sponsors of P4P initiatives include employer groups, health plans, provider organizations, Medicaid programs, and CMS.

In general, the literature finds that P4P initiatives are associated with better quality (Christianson, Leatherman and Sutherland, 2008a). But, it is difficult to determine if the initiatives caused the observed quality improvements, in part because financial incentives usually are employed as one component of an overall quality improvement strategy. Therefore, any quality gains that are observed could be due to payments, or they could be due to other features of the quality improvement strategy, or both.

There is no consistency in the design and implementation of the P4P initiatives that have been evaluated in the published literature. However, irrespective of the problem of determining the causal relationship between P4P initiatives and quality improvement in "real world" situations, the improvements reported in these evaluations are typically modest. Where quality gains are reported, they most often pertain to a subset of the measures on which payment was based (Christianson, Leatherman and and Sutherland, 2008b). Various explanations have been offered for the absence of major, significant effects associated with P4P initiatives, including:

- The incentive payments have not been large enough to change provider behavior.

- The existence of multiple different P4P initiatives sends confusing signals to providers.
- The wrong measures are being use to assess quality improvement.
- Providers do not have sufficient resources to invest in the major practice changes needed to improve quality.
- Providers face other, more pressing, issues (e.g. installing electronic medical records) that draw their attention away from responding to P4P initiatives (which could be considered a variant of "the payment is not large enough" argument.)
- The current practice environment for providers (especially physicians) leaves little time in a patient visit to take the necessary steps required to achieve a P4P reward; under fee-for-service payment, seeing more patients may be a more "efficient" way to increase practice income (if that is the goal).
- In some P4P schemes, the structure of payment creates uncertainty, which reduces the likelihood that providers will invest in practice changes necessary to receive a P4P award. For instance, if the reward is given to the top X% of providers, some providers may achieve a high level of performance on a P4P metric but may not be rewarded for their efforts.
- Physicians believe P4P programs encourage "check box" medicine and distract them from treating the "whole patient."

Most of these explanations relate to the design and implementation of P4P, not flaws in the basic concept. (Although, some analysts argue that incentive schemes such as P4P have not been particularly successful in non-health care settings in part because they are based on a flawed premise: that individual behavior is motivated primarily by financial incentives as opposed to a variety of other possible influences. For a general discussion, see Gagne and Deci, 2005) Because market conditions and the preferences of providers vary across locations and over time, there is no single, optimal P4P program structure. Instead, the challenge for P4P implementers is to design and carry out a P4P initiative that best fits their situation. This involves carefully considering different design options.

In this document, we identify the decisions that we believe, based on our review of the literature, are likely to be the most important in constructing a P4P initiative. However, the literature very seldom provides clear guidance concerning which decision to make, under which conditions. Among other things, this reflects the relatively small number of published evaluations of "real world" P4P programs from which to draw lessons, and the fact that (as noted above) P4P programs typically are embedded in broader quality improvement strategies,

which both affects their impact and makes that impact difficult to determine. Consequently, many aspects of P4P have been the subject of debate, without the development of a clear consensus (e.g see Fisher, 2006; Nelson, 2007).

Structure of Payment

A major issue in the design of any P4P program concerns whether to reward achievement of a predetermined benchmark of performance (e.g. 90% of patients in the practice with illness X receive test in time period T), or to reward a predetermined level of improvement (e.g. achievement of Z percent increase in the proportion of patients in the practice with illness X receiving test, measured from period T to period T+1.) There are other variations to consider. For instance, rather than rewarding improvement in the percent of patients receiving a test, if the measure is continuous the reward could be based on average amount of improvement (say, in a biologic marker.) A very different approach is to reward the top X percent of performers, using a predetermined formula to divide reward dollars (e.g. the top 5% of performers receive 50 % of the reward dollars, the next 5% receive 25%, and so forth.) This is called a "tournament" approach to compensation, because providers compete against each other for an award. The tournament approach can be applied to improvement as well. A final issue in the structure of payment relates to the use of penalties for low performers who do not improve over period T.

Rewarding the achievement of benchmarks: A major argument in favor of awarding P4P dollars for the achievement of pre-specified benchmarks is that the award process is easy to understand. The target is clear to providers, which should make their planning processes easier. For instance, they know that if they institute a new patient tracking process, and it is effective, they will receive a reward. In contrast, under a tournament approach, the provider may make the investment, perform at a higher level of quality as a result, but receive no reward if other providers perform better. From the provider's perspective, the uncertainty of obtaining the reward could discourage investment in quality improvement processes. However, there are three significant drawbacks associated with using pre-determined benchmarks to distribute award dollars. First, providers who already meet the benchmarks are essentially rewarded for their historical performance. They have little incentive to improve. This may not be a satisfactory result from the standpoint of payers who want their P4P programs to improve quality. Second, depending on where the benchmarks are set, providers who are at the low tail of the performance curve may have little incentive to invest in quality improvement. This would be the case if these providers did not believe they could improve enough to meet the benchmark in any given period. Conceptually, the use of benchmarks is likely to be the most effective if they are set at a level above current performance and if most providers felt that the

benchmarks could be achieved through reasonable efforts. Third, benchmarks can be problematic for payers because they complicate budgeting for the P4P initiative. If more providers achieve the benchmarks than predicted, costs could be greater than expected or budgeted. This risk increases in situations where there is not a good historical record of provider performance relative to the benchmark. Then, if the benchmark is set too low, a substantial portion of providers may have met it prior to implementation of the initiative. This problem can be solved by using a tournament approach, where the amount of reward dollars available is determined prior to implementing the P4P initiative. Then there is no risk that the payer will exceed the budgeted P4P amount.

Rewarding improvement: The alternative to using benchmarks for structuring rewards is to allocate award dollars based on percentage or nominal improvement. This has the advantage of providing incentives for low performing providers to invest in quality improvement. In fact, these providers may have a better chance of obtaining rewards than higher quality providers. This may be regarded as desirable if payers place greater weight on raising the quality of care offered by low-performing providers. However, it may not seem fair to other providers who perform better, but receive no award dollars, because their improvement was not as great. This issue is particularly relevant when, for the measures chosen, some providers already are performing near the maximum and have little or no possibility of earning an award.

Practice Example

There are two striking examples in the literature that highlight the challenges involved in structuring payments using a benchmark approach. In the United States, Rosenthal, et. al. (2005) evaluated a physician P4P program implemented by Pacificare, an HMO serving the western part of the United States (and subsequently acquired by UnitedHealthcare.) Pacificare rewarded performance on 5 ambulatory care quality measures and 5 patient-centered measures of service quality. Benchmarks were established based on the 75th percentile of the 2002 performance of the physician groups. These benchmarks were known to the participating medical groups, which also knew their own performance relative to the benchmarks prior to the initiative. An average medical group with 10,000 Pacificare patients had the potential to earn $270,000 per year, which was equal to 5% of the capitation payment to the group from Pacificare, but under 1% of an average group's total revenues. The program awarded about $3,4 million in bonuses from July, 2003-April, 2004 (27% of the potential bonus payments), and the evaluators found a significant, but modest, quality improvement in one of the three clinical quality measures they examined. Interestingly, according to the evaluators, "Physician groups whose performance payments were above the benchmark at baseline captured 75% of bonus payments" (p. 1792) for the measures they examined. In effect, three quarters of the award money was used to reward past performance of medical groups. However, in a somewhat

unexpected finding, groups that had the poorest quality scores at the beginning of the P4P initiative demonstrated the greatest improvement.

The second example concerns the physician P4P initiative in the U.K. (Roland, 2004). This program committed up to $3.2 billion in new funds over three years to reward general practitioners for performance relative to 146 quality indicators. Physicians earned points for percentages of patients in their practices meeting predetermined benchmarks. A 75% achievement of benchmarks overall was predicted when setting the budget for the initiative, but in the first year physicians achieved 96.7% of the points for quality indicators (Doran, et. al., 2006). Most of the budgeted monies for this three year program were dispersed in the program's first year. As a result, Campbell, et. al. (2007) observed that "The size of the gains in quality in relation to the costs of pay for performance remains a political issue in the United Kingdom, and the government now accepts that it paid more than it had expected to pay for the improvements in performance" (p. 189). In fact, the government did not have reliable data to use in setting the benchmarks and apparently underestimated existing quality levels. It is likely that a significant portion of the funds actually rewarded practices for their historical performance. Because there were not good baseline measures of quality, it is unclear whether or not the P4P program actually resulted in significant quality improvements in primary care in the U.K.

Amount of Payment

A second major issue relates to the amount of payment necessary to achieve improvement. This can be thought of in terms of a specific dollar amount or a percent of practice income. The "right amount" necessary to achieve the desired result likely varies with the structure of the payment and how the payment, if received at the physician group level, is used (e.g. included in incentive payments to physicians or physician practices or used of improving group infrastructure, etc.)

Practice Example

There is almost no research that addresses the level of payment needed to achieve desired results in a P4P program. To do so would require a study design in which several P4P programs were compared that were identical except for differences in reward levels. The closest that any published study comes to this design is the comparative analysis of five Medicaid programs conducted by Felt-Lisk, et. al., (2007). However, because the Medicaid programs varied along dimensions other than size of payment, this evaluation could conclude only that the greatest response occurred in the program that offered the largest rewards.

The literature does highlight the considerable variation in potential (and, in some cases, actual) rewards to be found in P4P programs. The U.K. P4P program offered the most generous

rewards found in any of the P4P programs where evaluations have been published. At its inception, that program offered the potential for physicians to increase their practice incomes by $77,000 per physician (Roland, 2004); in practice during its first year the program increased the income of general practitioners by an average of $40,000, a considerable percentage increase over their average income of $122,000 to $131,000 before the program was put in place (Doran, et. al., 2006). As noted above, the potential increase in income under Pacificare's P4P program was less than 1% of a medical group's annual revenues.

Most published evaluations do not provide detailed information on the size of the reward in the P4P program, either in absolute or in relative terms. Where this information is provided, the reward typically is relatively small. For instance, the maximum reward for top-performing hospitals in the CMS/Premier P4P demonstration was a 2% increase in Medicare reimbursements, which is an increase of 1% or less in total revenues for most hospitals (Lindenauer, et. al., 2007). In early experiments conducted by Hillman, et. al. (1998, 1999), the bonus for top performing physician practices was 10 percent of a practice's typical capitation payment from the payer. In commenting on the incentive's lack of impact, the authors note that this payer was one of many for these practices. Kouides et. al. (1998) evaluated the impact of paying physicians $0.80 per influenza immunization if their practice immunization rates exceeded 70% and $1.60 if they exceeded 85%. Immunization rates improved by a greater amount in the incentive group, relative to a group of physicians that did not receive incentives, even though Kouides, et. al. (1998) characterized the incentives as "modest." In a P4P program implemented by a health plan in Hawaii, physicians received an average bonus payment of 3.5% for attaining predetermined benchmarks, and physicians who showed significant improvement in scores received a bonus payment of $3,000. In a health plan-sponsored program aimed at diabetes care improvement, the annual distribution of payments ranged from $6,000 to $18,000 (Curtin, et. al., 2006). Larson, Cannon, and Towner (2003) report a relatively small incentive of from 0.5 to 1 per cent of physician income for improvement of care along several dimensions, with half of the payment going towards rewarding improvements in diabetes care.

As these examples illustrate, the level of the awards found in P4P programs varies enormously. In some cases, it appears that reward amounts were set in negotiations with participating physicians, and in other cases they seem to have been set unilaterally by payers. Irrespective of how they were set, "real world" P4P programs are not structured in a way that allows researchers to answer a question that is of fundamental interest to payers: "how much is enough" to generate significant improvements in quality.

Type of Measure

A third design feature concerns the types of measures to use when rewarding performance. The basic issue is whether to use "process of care" measures or "outcome"

measures, or some combination of the two. The choice has implications for the cost of data collection and measure construction (and, therefore, the cost-effectiveness of P4P). There are varying arguments. For instance, treatment processes are under the control of providers, whereas outcome measures may be influenced by patient behaviors that providers may not be able to control. Thus, from a provider perspective, process measures may be regarded as a "fairer" basis for rewarding provider performance. But, from a payer perspective, the goal of a P4P initiative ultimately is to improve patient health, so payers may favor P4P programs based on outcome measures.

Practice Example

Most measures used in P4P programs sponsored by health plans in the United States are constructed using claims data (e.g. see Chung, et. al., 2003), and therefore they are process of care measures. Generally, they try to capture the conformance of care with widely-accepted evidence-based treatment guidelines (e.g. see Felt-Lisk, et. al., 2007; Greene, et. al., 2004). They tend to focus on preventive actions relating to screening (e.g. see Armour, et. al., 2004; Langham, Gillam, and Thorogood, 1995; Rosenthal, et. al., 2005) and receipt of immunizations (e.g. Morrow, Gooding and Clark, 1995), and on treatment of chronic illnesses where there are widely accepted medical standards. More indicators are available for diabetes that for any other chronic illness (e.g. see Beaulieu and Horrigan, 2005; Curtin, et. al., 2006; Young, et. al., 2007). Treatment of heart conditions is another area of care that receives considerable attention in P4P initiatives (e.g. see Glickman, et. al., 2007; Nahra, et. al., 2006). Some health plans have included measures that go beyond items for which claims are processed (e.g. delivery of smoking cessation advice). These measures typically require documentation in the medical records of patients, and performance is assessed based on a random sample of patient records. Some P4P programs use a variety of different measures, with the P4P program in the U.K. providing the best example of this. In this program, rewards are given for certain practice characteristics and for patient ratings, along with standard process of care indicators. In general, payers appear to be expanding the number of outcome measures used in their P4P programs, especially intermediate outcome measures such as lipid levels, blood pressure readings and HbA1c levels. These data are available in patient medical records and typically are collected and submitted by the physician practice. The increased use of electronic medical records facilitates this approach, but constructing these measures remains more expensive for providers, and for payers as well if the data are audited. Measures based on patient reports are the most expensive to construct and also are the least common types of measures presently used in P4P programs in the United States. Five of the 10 measures used by Pacificare in its P4P program (see above) were patient-reported measures of service quality (Rosenthal, et. al., 2005)

Number of Measures

Determining the number of measures to use, irrespective of the type of measure, is an important design decision in any P4P program. The argument for using a large number of measures is that it encourages overall improvement in quality. The argument against it is that the incentive to improve in any one area is weak. Advocates of fewer measures say this can focus provider resources on areas where improvement is needed the most. However, a contrary view is that providers may focus too strongly on the targets, and quality may decrease in areas not included in the P4P program. This concern is no different from that sometimes expressed about paying teachers or schools based on student performance on specific standardized tests. The issue is whether, in the presence of rewards, teachers will "teach to the test;" that is, focus on subjects that will be covered in the test, to the detriment of student learning in other areas.

Practice example

Active P4P programs vary widely in the number of P4P measures they use. For example, the P4P program aimed at general practitioners in the U.K uses 146 quality indicators while programs sponsored by health plans in the U.S. typically use 10 or fewer measures designed to reflect well-established best practices in preventive care and in the treatment of some chronic illnesses. For instance, researchers have evaluated the impact of several different P4P initiatives addressing only diabetes care. Research relating to the number of measures used in P4P programs has focused primarily on documenting the impact of P4P on un-rewarded aspects of quality.

Glickman, et al. (2007) assessed the impact of the CMS/Premier hospital P4P initiative (described above) on AMI process of care measures. Six measures were included in the P4P program, and the authors tracked these measures, as well as 8 other AMI treatment quality indicators not rewarded under the program. They found significant improvements in two P4P measures where no change in physician practice was required, and the cost of change for hospitals was relatively low (aspirin at discharge and smoking cessation counseling), but no improvement in a composite performance measure that included all 6 P4P metrics. The authors also found no effect—negative or positive—on any of the measures not included in the CMS/P4P initiative. Beaulieu and Horrigan (2005) evaluated a physician P4P program for diabetes care sponsored by a health plan in the United States. Based on data from a small number of physician practices, they reported improvement in 5 of 6 diabetes measures in the P4P program and no impact on quality of care in areas not targeted by P4P for rewards. In evaluating the impact of the U.K.'s P4P program on 42 primary care practices, Campbell, et. al. (2007) focused on measures of quality for heart disease, asthma and type 2 diabetes. They

found significant, but modest, improvements in asthma and diabetes quality of care indicators, and no negative impacts on other measures of quality not incented by the P4P program.

In summary, concerns that P4P programs could have a negative impact on quality in areas of care where financial incentives were not applied are not supported by existing research. There are (at least) two explanations for this research finding. First, there is little credible evidence in evaluations of P4P programs to date that these programs have resulted in substantial improvements in the indicators that they have targeted with financial incentives. It would seem most reasonable to expect negative impacts on other areas of quality in situations where P4P programs had an impact on their targeted measures. Second, it may be that significant resource shifts within practices are not necessary to secure P4P rewards related to many quality measures, so that non-targeted areas of care are not put "at risk" by P4P programs.

Aggregation of Measures

P4P programs that use multiple measures of performance must decide how they will use information from these measures to pay providers. The basic choice is whether to pay providers based on their scores on an aggregate measure of performance or to pay separately for performance on individual measures, with the total payment equal to the sum of these individual payments. When paying providers based on an aggregate score, P4P implementers must decide how to combine the individual scores; that is, they must decide the weight to give to each component in creating the aggregate measure. We could find no discussion of the strengths and limitations of these different approaches in the literature. It seems reasonable to propose that, if an aggregate measure is constructed, the aggregation process should be transparent to providers participating in the P4P program. Providers seeking to achieve rewards for improving quality or achieving threshold levels of quality should have a clear understanding of which aspects of quality are valued the most by the payer. It also is important for payers to understand that a simple adding up of achievement on different measures to form the aggregate score implicitly means that all improvements are valued equally.

Practice Example

The Pacificare P4P program used 10 different measures of quality; 5 are clinical measures and 5 are measures of patient satisfaction. Performance on each measure was rewarded separately, so that the medical group's total reward was the sum of these individual rewards. In contrast, in the CMS/Premier hospital demonstration payment to hospitals was made based on an aggregate measure. According to Lindenauer, et. al (2007), "For each of the clinical conditions, hospitals performing in the top decile on a composite measure of quality for a given year received a 2% bonus payment in addition to the usual Medicare reimbursement

rate" (p. 488). The Premier website describes the aggregation process in detail, with examples. Hospitals begin by submitting their raw data to Premier, which calculates a quality index for each clinical area included in the program. The overall index consists of a "Composite Process Rate" and a "Risk-Adjusted Outcomes Index." To calculate the Composite Process Rate, the numerator values for each of the process measures are summed to create a composite numerator, with a composite denominator calculated in the same way. The composite numerator is divided by the composite denominator to generate the overall Composite Process Rate. For outcomes measures, a hospital's actual outcomes rate is divided by its risk-adjusted rate and the result is multiplied by 100. A final Composite Quality Index in each area is calculated by weighting each score within each area equally. The scores are ranked, and hospitals in the top decile or second decile receive a predetermined payment for their performance in that area. This process is repeated for every clinical area included in the program.

In the U.K P4P initiative, a different number of points was awarded for different performance indicators. Martin Roland, MD, one of the architects of the U.K.'s program, states that "Family practitioners can now earn up to 1000 'points' for achievement in relation to the complex set of indicators that make up the Quality and Outcomes Framework" (Roland, 2004). As an example, for patients with heart disease, if blood pressure has been recorded in the previous 15 months for 25% of patients, the practice receives 1 point. If it has been recorded for 90% or more of patients, the practice receives 7 points. Points for all clinical performance areas are awarded in this same general way. With respect to practice organization, points are often awarded for the presence or absence of some desirable feature; for example, a practice can receive 1.5 points if there are "clearly defined arrangements for backing up computer data" (Roland, 2004, p. 1451). After all points have been determined and summed, the result is multiplied by a predetermined "per point" amount to calculate the total payment to the practice.

We found no studies in the literature that investigated whether paying for performance based on an aggregate measure, or rewarding individual measures, leads to better performance.

<u>Determining the Denominator for Constucting Measures</u>

Which patients should be included when constructing performance measures to use in awarding P4P payments? If patients are included inappropriately, the provider has a financial incentive to give care that may not be needed, which could increase costs and raise issues of patient safety. For example, a common process is to use all patients meeting specific diagnoses criteria (e.g. to establish that the patient has diabetes) in calculating performance on a specific measure. But, this may not always make sense; a physician treating a diabetic patient with

terminal cancer arguably should not be penalized if the patient does not receive a scheduled foot exam according to guidelines. In designing P4P initiatives, there are essentially three approaches to addressing this problem. First, threshold performance levels may be set at some target less that 100%. For instance, an 80% threshold would not penalize a provider for using her best clinical judgment in not providing guideline-recommended treatments to up to 20% of her patients. The drawback of this approach is that, if the threshold is set too low, achieving a level of performance necessary to receive a P4P reward may be "too easy," and there may be little actual quality improvement as a result. Second, providers can be allowed to formally exclude patients from measurement, if the patients meet predetermined criteria. The drawback of this approach is that the criteria for exclusion may be too general, creating considerable latitude for the provider to construct the panel of patients to be used for measurement purposes. The result could be provider "gaming" of the process, inappropriately excluding patients who would bring down the average performance of the practice. Third, a statistical risk-adjustment technique can be applied to "level the playing field" among providers in the P4P program. If the risk adjustment approach is effective, the performance of providers is compared for an "average" panel of patients. The drawback of using a statistical risk adjustment methodology is that it may not be transparent to providers, and may cause confusion and suspicion. Also, statistical risk–adjustment methodologies may not remove enough of the performance variation associated with patient characteristics to adequately address the denominator problem.

Determining the appropriate denominator for measurement purposes is an important design decision because it can have a major impact on the money paid out under P4P programs, as well as which providers receive payments. For example, research suggests that excluding as few as three diabetic patients in a primary care physician's practice can have a large impact on average practice performance (Hofer, et. al, 1999.) And, done incorrectly, it can raise questions about the validity of the entire methodology for allocating reward dollars.

Practice Example

Evaluators of the U.K. P4P initiative have explored issues that can arise in determining the denominator for measuring performance. In this program, physicians were able to use various criteria to exclude individual patients from calculations of quality measures, a process called "exception reporting." The following reasons are permitted for exclusion of patients from the measurement denominator used to reward GP performance (Doran, 2008, p. 276):

1. "The patient has received at least three invitations for review during the preceding 12 months but has not attended."

2. "The indicator is judged to be inappropriate for the patient because of particular circumstances, such as terminal illness, extreme frailty, or the presence of a supervening condition that makes the specified treatment clinically inappropriate."

3. "The patient has recently received a diagnosis or has recently registered with the practice."

4. "The patient is taking the maximum tolerated dose of a medication, but the levels remain suboptimal."

5. "The patient has had an allergic or other adverse reaction to a specified medication, but the levels remain suboptimal."

6. "The patient does not agree to the investigation or treatment."

7. "A specified investigative service is unavailable to the family practitioner."

Evidence from the first year of the U.K.'s P4P initiative suggested that "gaming" of the exclusion criteria might have been an issue. The factor that had the greatest impact in explaining variation on performance across practices was exception reporting; an increase in 1 percent in the proportion of patients excluded was associated with an increase of .31% in performance. Overall rates of exception reporting ranged from 0 to 85%, suggesting that at least some practices may have engaged in excessive exception reporting (Doran, et. al., 2006). However, more recent work with better data suggests these fears likely were unfounded. The median percent of patients excluded in the second year of the program was 5.3%, with characteristics of physicians and patients explaining less than 3% of the variation in exception reporting. The authors estimate that exception reporting "accounted for approximately 1.5% of the cost of the pay-for-performance" program in that year (Doran, et. al., 2008, p. 274).

The U.K. experience with a relatively simple and transparent approach to "leveling the playing field" when measuring performance is promising. Assuring a "level playing field" becomes especially desirable when outcome measures are used in pay-for-performance programs, because then patient behaviors and characteristics are likely to play even more important roles in achieving goals and determining payouts.

Adjusting for Practice Characteristics

It is frequently argued that some practices are better able to respond to P4P programs because they 1) have more resources, financial and otherwise, or 2) their patients are better able to adhere to treatment plans and carry out self-management activities. The latter issue can be addressed, at least conceptually, through risk adjustment techniques that incorporate patient socio-demographic characteristics. However, the absence of a "level playing field" in terms of practice resources is more difficult to address. Potentially, it could have longer-term

consequences for the quality of health care received by economically disadvantaged groups of patients.

The concern is that, if disadvantaged patients make up a disproportionately large percentage of patients in some practices, then these practices may receive no payments (if the patients are uninsured) or low payments (if the patients are enrolled in Medicaid) for providing services to many of their patients. These practices then would be less likely to have the financial resources required to make the investments needed to achieve P4P benchmarks and receive P4P awards. If this is the case, the difference in the financial condition of practices serving significant numbers of disadvantaged patients and other practices could widen. That is, P4P could contribute to a situation where the "rich get richer and the poor get poorer." If this occurs, P4P also could contribute to widening existing racial and ethnic disparities in quality of health care.

There are two ways to address this concern. First, as already mentioned, risk-adjustment approaches could be employed, so that awards were based on an "average" patient panel. Second, dollars could be allocated directly to these practices to improve their infrastructures where this seemed warranted.

Practice Example

In the U.K., concern about the ability of practices located in lower income areas to compete effectively for P4P rewards were expressed prior to program implementation. Consequently, researchers in the U.K. looked for evidence of any problems in this regard. In particular, Doran, et al. (2008) examined the relationship between degree of "deprivation" in the census areas in which physician practices were located and performance on quality indicators over the first three years of the program. They found that median achievement levels on the indicators grew for practices in both the lowest and highest income areas and that the gap in achievement between these practices actually narrowed over time. This suggests that, contrary to concerns, the practices in low income areas were not hurt by the P4P program and their patients benefited from improvement in the quality of care they received. However, as described above, the U.K had an "exception" system in place that may have contributed to this favorable finding.

Also in the U.K., Guilford, et al. (2007) examined the same issue, with a focus on the performance of practices in serving diabetes patients. They found that practices located in "deprived" areas were less successful than other practices in achieving P4P benchmarks in the first year of the P4P programs. This is a different finding than reported by Doran, et. al. (2007) but the two sets of results are not necessarily in conflict; Doran, et. al. (2008) examined data over a longer time period and for more measures.

Srirangalingam, et. al. (2006) conducted an analysis of treatment for diabetes in clinics located in deprived areas of central London, examining how referral patterns changed after implementation of the P4P initiatives. They reported a significant increase in referrals to specialists for patients with poor blood sugar control. It is not clear if this resulted in better quality of care for these patients. Also, the referrals may have changed the panel of patients employed to measure performance in GP practices, improving the probability of practices receiving a P4P award.

There has been much less attention devoted to this issue by program evaluators in the United States, possibly because most "real world" P4P programs have been implemented by health plans and therefore have affected primarily privately insured

populations. Several early experimental studies of P4P in the U.S. were carried out in Medicaid environments, but both experimental and control practices served large numbers of Medicaid recipients. Therefore, there was little opportunity to compare the results for practices serving low income populations to other practices. Karve, et. al. (2008) estimated a statistical relationship between a hospital's performance in Medicare's P4P program and the proportion of patients who were African American. They found that having a higher proportion of African American patients was associated with lower levels of performance on P4P indicators related to treatment for AMI and community-acquired pneumonia.

Based on the existing research, the impact of P4P on practices serving disadvantaged populations, as well as its ultimate impact on racial disparities in quality of care and health outcomes, is unclear. This is a question that deserves greater research attention as P4P initiatives expand to encompass all population segments in communities.

REFERENCES

1. Achat H, McIntyre P and Burgress M (1999). 'Health care incentives in immunization'. *Australian and New Zealand Journal of Public Health,* vol 23, pp 285–88.

2. Amundson G, Solberg LI, Reed M, Martini EM and Carlson R (2003). 'Paying for quality improvement: compliance with tobacco cessation guidelines'. *Joint Commission Journal on Quality & Safety*, vol 29, pp 59–65.

3. Armour BS, Friedman C, Pitts M, Wike J, Alley L, and Etchason J (2004). 'The influence of year-end bonuses on colorectal cancer screening'. *The American Journal of Managed Care* vol 10, pp. 617-24.

4. Armour BS, Pitts MM, Maclean R, Cangialose C, Kishel M, Imai H and Etchason J (2001). 'The effect of explicit financial incentives on physician behavior'. *Archives of Internal Medicine*, vol 161, pp 1261–66.

5. Ashworth M, Armstrong D, de Freitas J, Boullier G, Garforth J and Virji A (2005). 'The relationship between income and performance indicators in general practice: a cross-sectional study'. *Health Services Management Research*, vol 18, pp 258–64.

6. Beaulieu ND and Horrigan DR (2005). 'Organisational processes and quality: putting smart money to work for quality improvement'. *Health Services Research,* vol 40, 1318–34.

7. Bhattacharyya T, Mehta P, and Freiberg AA (2008). 'Hospital characteristics associated with success in a pay-for-performance program in orthopaedic surgery'. *The Journal of Bone & Joint Surgery* vol 90, pp. 1240-43.

8. Cameron PA, Kennedy MP and McNeil JJ (1999). 'The effects of bonus payments on emergency service performance in Victoria'. *The Medical Journal of Australia*, vol 171, pp 243–46.

9. Campbell SM, McDonald R, and Lester H (2008). 'The experience of pay for performance in English family practice: a qualitative study'. *Annals of Family Medicine* vol 6, pp. 228-34.

10. Campbell S, Reeves D, Kontopantelis E, Middleton E, Sibbald B and Roland M (2007). 'Quality of primary care in England with the introduction of pay for performance'. *The New England Journal of Medicine*, vol 357, pp 181–90.

11. Casale AS, Paulus RA, Selna MJ, Doll MC, Bothe Jr AE, McKinley KE, Berry SA, Davis DE, Gilfillan RJ, Hamory BH, and Steele Jr GD (2007). '"ProvenCare℠" A provider-driven pay-for-performance program for acute episodic cardiac surgical care'. *Annals of Surgery* vol 246, pp. 613-23.

12. Chiang C-Y, Enarson DA, Yang S-L, Suo J, Lin T-P (2002). 'The impact of national health insurance on the notification of tuberculosis in Taiwan'. *The International Journal of Tuberculosis and Lung Disease*, vol 6, pp 974–79.

13. Christensen DB, Neil N, Fassett WE, Smith DH, Holmes G and Stergachis A (2000). 'Frequency and characteristics of cognitive services provided in response to a financial incentive'. *Journal of the American Pharmaceutical Association,* vol 40, pp 609–17.

14. Christianson JB, Leatherman S, and Sutherland K (2008a). 'Lessons from evaluations of purchaser pay-for-performance programs'. *Medical Care Research and Review*, vol 65, pp 5S-35S.

15. Christianson JB, Leatherman S, and Sutherland K (2008b). *Financial Incentives, Healthcare Providers and Quality Improvements. A Review of the Evidence*. London, England: The Health Foundation.

16. Christianson JB, Leatherman S, and Sutherland K (2007). 'Paying for quality: Understanding and assessing physician pay-for-performance initiatives'. RWJF Research Synthesis Report No. 13. Princeton, NJ: The Robert Wood Johnson Foundation.

17. Chung RS, Chernicoff HO, Nakao KA, Nickel RC and Legorreta AP (2003). 'A quality-driven physician compensation model: Four-year follow-up study'. *Journal for Healthcare Quality*, vol 25, pp 31–37.

18. Collier VU (2007). 'Use of pay for performance in a community hospital private hospitalist group: a preliminary report'. *Transactions of the American Clinical and Climatological Association* vol 118, pp. 263-72.

19. Curtin K, Beckman H, Pankow G, Milillo Y and Greene RA (2006). 'Return on investment in pay for performance: a diabetes case study'. *Journal of Healthcare Management*, vol 51, pp 365–76.

20. Damberg CL, Raube K, Williams T and Shortell SM (2005). 'Paying for performance: implementing a statewide project in California'. *Quality Management in Health Care*, vol 14, pp 66–79.

21. Doran T. and Fullwood C (2007). 'Pay for performance: is it the best way to improve control of hypertension'? *Current Hypertension Reports* vol 9, pp. 360-67.

22. Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U and Roland M (2006). 'Pay-for-performance programmes in family practices in the United Kingdom'. *The New England Journal of Medicine*, vol 335, pp 375–84.

23. Doran T, Fullwood C, Kontopantelis E, and Reeves D (2008). 'Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework'. *Lancet* vol 372, pp. 728-36.

24. Doran T, Fullwood C, Reeves D, Gravelle H, and Roland M (2008). 'Exclusion of patients from pay-for-performance targets by English physicians'. *The New England Journal of Medicine* vol 359, pp. 274-84.

25. Dudley RA, Frolich A, Robinowitz DL, Talavera JA, Broadhead P, Luft HS and McDonald K (2004). *Strategies to Support Quality-Based Purchasing: A review of the evidence.* Technical Review Number 10. AHRQ Publication No. 04-0057, Agency for Healthcare Research and Quality.

26. Ettner SL, Thompson TJ, Stevens MR, Mangione CM, Kim C, Steers WN, Goewey J, Brown AF, Chung RS, Vankat Narayan KM and the TRIAD Study Group (2006). 'Are physician reimbursement strategies associated with processes of care and patient satisfaction for patients with diabetes in managed care?' *Health Services Research*, vol 41, pp 1221–41.

27. Fairbrother G, Hanson KL, Friedman S and Butts GC (1999). 'The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates'. *American Journal of Public Health*, vol 89, pp 171–75.

28. Fairbrother G, Siegel MJ, Friedman S, Kory PD and Butts GC (2001). 'Impact of financial incentives on documented immunization rates in the inner city: results of a randomised controlled trial'. *Ambulatory Pediatrics*, vol 1, pp 206–12.

29. Felt-Lisk S, Gimm G and Peterson S (2007). 'Making pay-for-performance work in Medicaid'. *Health Affairs – Web Exclusive,* June 26, w516–w527.

30. Fisher ES (2006). 'Paying for performance: risks and recommendations'. *The New England Journal of Medicine*, vol 355, pp 1845–47.

31. Gagné M and Deci EL (2005). 'Self-determination theory and work motivation'. *Journal of Organisational Behaviour*, vol 26, pp 331–62.

32. Gené-Badia J, Escaramis-Babiano G, Sans-Corrales M, Sampietro-Colom L, Aguado-Menguy F, Cabezas-Peña C and Gallo de Puelles P (2007). 'Impact of economic incentives on quality

of professional life and on end-user satisfaction in primary care'. *Health Policy*, vol 80, pp 2–10.

33. Glickman SW, Ou F-S, DeLong ER, Roe MT, Lytle BL, Mulgund J, Rumsfeld JS, Gibler WB, Ohman EM, Schulman KA and Peterson ED (2007). 'Pay for performance, quality of care, and outcomes in acute myocardial infarction'. *Journal of the American Medical Association*, vol 297, pp 2373–80.

34. Grady KE, Lemkau JP, Lee NR and Caddell C (1997). 'Enhancing mammography referral in primary care'. *Preventive Medicine*, vol 26, pp 791–800.

35. Greene RA, Beckman H, Chamberlain J, Partridge G, Miller M, Burden D and Kerr J (2004). '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, pp 670–78.

36. Grossbart SR (2006). 'What's the return? Assessing the effect of 'pay-for-performance' initiatives on the quality of care delivery'. *Medical Care Research & Review*, vol 63 (1 Suppl), pp 29S–48S.

37. Gulliford MC, Ashworth M, Robotham D and Mohiddin A (2007). 'Achievement of metabolic targets for diabetes by English primary care practices under a new system of incentives'. *Diabetic Medicine*, vol 24, pp. 505-11.

38. Hillman AL, Ripley K, Goldfarb N, Nuamah I, Weiner J and Lusk E (1998). 'Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care'. *American Journal of Public Health*, vol 88, pp 1699–1701.

39. Hillman AL, Ripley K, Goldfarb N, Weiner J, Nuamah I and Lusk E (1999). 'The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care'. *Pediatrics*, vol 104, pp 931–35.

40. Hofer TG, Hayward RA, Greenfield S, Wagner EH, Kaplan SH and Manning WG (1999). '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, pp 2098–105.

41. Karve A, Ou F-S, Lytle BL, and Peterson ED (2008). 'Potential unintended financial consequences of pay-for-performance on the quality of care for minority patients.' *American Heart Journal* vol 155, pp. 571-76.

42. Khunti K, Gadsby R, Millett C, Majeed A and Davies M (2007). 'Quality of diabetes care in the U.K.: comparison of published quality-of-care reports with results of the quality and outcomes framework for diabetes'. *Diabetic Medicine*, vol 24, pp. 1436-41.

43. Kouides RW, Bennett NM, Lewis B, Cappuccio JD, Barker WH and LaForce FM (1998). '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, pp 89–95.

44. Kraft AD, Capuno JJ, Quimbo SA, and Tan Jr CAR (2008). 'Information, incentives and practice patterns: the case of TB dots services and private physicians in the Philippines'. *The Singapore Economic Review* vol 53, pp. 43-56.

45. Langham S, Gillam S and Thorogood M (1995). 'The carrot, the stick and the general practitioner: how have changes in financial incentives affected health promotion activity in general practice?' *British Journal of General Practice*, vol 45, pp 665–68.

46. Larsen DL, Cannon W and Towner S (2003). 'Longitudinal assessment of a diabetes care management system in an integrated health network'. *Journal of Managed Care Pharmacy*, vol 9, pp 552–58.

47. Levin-Scherz J, DeVita N and Timbie J (2006). 'Impact of pay-for-performance contracts and network registry on diabetes and asthma HEDIS measures in an integrated delivery network'. *Medical Care Research & Review*, vol 63(1 Suppl), pp 14S–28S.

48. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A and Bratzler DW (2007). 'Public reporting and pay for performance in hospital quality improvement'. *The New England Journal of Medicine*, vol 356, pp 486–96.

49. McDonald R, Harrison S, Checkland K, Campbell SM and Roland M (2007). 'Impact of financial incentives on clinical autonomy and internal motivation in primary care: ethnographic study'. *British Medical Journal*, vol 334, pp 1357–62.

50. McDonald R, Harrison S, and Checkland K (2008). 'Incentives and control in primary health care: findings from English pay-for-performance case studies'. *Journal of Health Organization and Management* vol 22, pp. 48-62.

51. Mehrotra A, Pearson SD, Coltin KL, Kleinman KP, Singer JA, Rabson B, and Schneider EC (2007). 'The response of physician groups to P4P incentives'. *The American Journal of Managed Care* vol 13, pp. 249-55.

52. Millett C, Gray J, Saxena S, Netuveli G, and Majeed A (2007). 'Impact of a pay-for-performance incentive on support for smoking cessation and on smoking prevalence among people with diabetes'. *Canadian Medical Association Journal* vol 176, pp. 1705-10.

53. Morrow RW, Gooding AD and Clark C (1995). 'Improving physicians' preventive health care behaviour through peer review and financial incentives'. *Archives of Family Medicine*, vol 4, pp 165–69.

54. Nahra TA, Reiter KL, Hirth RA, Shermer JE, Wheeler JRC (2006). 'Cost-effectiveness of hospital pay-for-performance incentives'. *Medical Care Research and Review*, vol 63, pp 49S–72S.

55. Nalli GA, Scanlon DP and Libby D (2007). 'Developing a performance-based incentive program for hospitals: a case study from Maine'. *Health Affairs*, vol 26, pp 817–24.

56. Nelson AR (2007). 'Pay-for-performance programs: ethical questions and unintended consequences'. *Current Clinical Practice*, vol 1, pp 16-8.

57. Parke II DW (2007). 'Impact of a pay-for-performance intervention: financial analysis of a pilot program implementation and implications for ophthalmology (an American Ophthalmological Society thesis)'. *Transactions of the American Ophthalmological Society* vol 105, pp. 448-60.

58. Petersen LA, Woodard LD, Urech T, Daw C and Sookanan S (2006). 'Does pay-for-performance improve the quality of health care?' *Annals of Internal Medicine*, vol 145, pp 265–72.

59. Ritchie LD, Bisset AF, Russell D, Leslie V and Thomson I (1992). 'Primary and preschool immunization in Grampian: progress and the 1990 contract'. *British Medical Journal*, vol 34, pp 816–19.

60. Roland M (2004). 'Linking physicians' pay to the quality of care: a major experiment in the United Kingdom'. *The New England Journal of Medicine*, vol 251, pp 1448–54.

61. Rosenthal MB and Frank RG (2006). 'What is the empirical basis for paying for quality in health care?' *Medical Care Research & Review*, vol 63, pp 135–57.

62. Rosenthal MB, Frank RG, Li Z and Epstein AM (2005). 'Early experience with pay-for-performance: from concept to practice'. *Journal of the American Medical Association*, vol 294, pp 1788–93.

63. Scott A and Hall J (1995). 'Evaluating the effects of GP remuneration: problems and prospects'. *Health Policy*, vol 21, pp 183–95.

64. Simpson CR, Hannaford PC, Lefevre K and Williams D (2006). 'Effect of the U.K. incentive-based contract on the management of patients with stroke in primary care'. *Stroke*, vol 37, pp 2354–60.

65. Srirangalingam U, Sahathevan SK, Lasker SS and Chowdhury TA (2006). 'Changing pattern of referral to a diabetes clinic following implementation of the new UK GP contract'. *British Journal of General Practice,* vol 56, pp 624–26.

66. St. Jacques PJ, Patel N and Higgins MS (2004). 'Improving anesthesiologist performance through profiling and incentives'. *Journal of Clinical Anesthesia,* vol 16, pp 523–28.

67. Sutton M and McLean G (2006). 'Determinants of primary medical care quality measured under the new UK contract: cross sectional study'. *British Medical Journal*, vol 332, pp 389–90.

68. Town R, Kane R, Johnson P and ButlerM (2005). 'Economic incentives and physicians' delivery of preventive care: a systematic review'. *American Journal of Preventive Medicine*, vol 28, pp 234–40.

69. Whalley D, Gravelle H, and Sibbald B (2008). 'Effect of the new contract on GPs' working lives and perceptions of quality of care: a longitudinal survey'. *British Journal of General Practice* vol 58, pp. 8-14.

70. Young GJ, Meterko M, Beckman H, Baker E, White B, Sautter KM, Greene R, Curtin K, Bokhour BG, Berlowitz D, and Burgess Jr JF (2007). 'Effects of paying physicians based on their relative performance for quality'. *Journal of General Internal Medicine* vol 22, pp. 872-76.

APPENDIX

Summaries of Published Studies*

SUMMARIES OF REVIEW ARTICLES

Achat, McIntyre and Burgress (1999) reviewed use of incentives to influence immunisation uptake, identified issues in developing incentive programmes and examined findings in the context of a new immunisation incentive scheme in Australia. They conducted a MEDLINE search, in English, under immunisation and financial incentives, from 1966 to 1998. They discussed a U.S. study in New York by Kouides *et al* (1998) which found that, when primary care physicians were rewarded for reaching a 70 per cent target with a fee increase of 10 per cent, the average rate was 73.1 per cent compared with 55.7 per cent in a comparison group. Incentives were less influential in practices with fewer than 100 patients. Ritchie *et al* (1992) looked at changes in rates after the implementation of a new contract for GPs in Scotland in 1990. GPs received additional payments of £1800 (high target – 90 per cent) and £600 (low target). The number of physicians achieving 95 per cent or more rose from 31 to 81 per cent for primary immunisation and from 23 to 64 per cent for preschool boosters. The reasons for the increase were not clear, and there were other factors at work in addition to the financial incentives. Based on the discussion provided, it is not possible to determine the strength of the study designs used by these authors.

Armour *et al* (2001) reviewed the impact of explicit financial incentives at the physician-level on resource use (hospital and visits) and quality measures. The literature review was conducted following the Cochrane Collaborative handbook. The authors did not state how many articles were identified through their review, but they discussed two articles related to resource use and four related to quality of care. One article related to resource use was based on data from a survey of medical directors (Hillman, 1989). The second examined the impact of bonus payments at the physician versus the physician group level. Incentives directed at the individual physician-level were found to be the most effective. The authors reported mixed results regarding the four studies where quality measures were used as outcome variables. One study found no impact while another reported that quality of care, measured by children's immunisation rates, improved. The authors noted the very limited amount of research related to the impact of imposing direct financial incentives on physicians.

* These summaries are part of a larger discussion prepared for The Health Foundation in the United Kingdom (see Christianson, Leatherman, and Sutherland, *Financial Incentives, Healthcare Providers and Quality Improvements. A Review of the Evidence*. London, England: The Health Foundation, 2007).

Dudley *et al* (2004) conducted a literature review of the evidence on strategies to support quality-based purchasing, which includes a review of the literature on use of financial incentives for providers to improve quality. The authors concluded that a performance-based provider payment could 'plausibly be introduced by a purchaser'. A variety of different outcomes were measured across the studies that were reviewed. The authors interrogated MEDLINE and Cochrane databases, as well as databases documenting ongoing work. Eight randomised trials in which the trial used a performance-based payment as the intervention were identified and included in the review. In four of the articles the recipient of the incentive payment was the individual provider, while in the other four the recipient was either a provider or provider group. In the four studies where the incentive targeted the individual provider, there were five positive and two negative results. In the remaining studies there were one positive and two negative results, where 'positive' indicates a result in the desired direction, and 'negative' means there was no significant effect. In seven studies the target for the incentive was the physician. In these studies there were five significant positive effects and four cases of no significant effect. Positive effects were more likely to be observed when the incentive took the form of an addition to fee-for-service payment than when the incentive was paid as a bonus. Seven studies (and nine dependent variables) addressed preventive care.

Petersen *et al* (2006) reviewed the literature on studies where there was an explicit financial incentive to improve quality. They conducted a PubMed search of the English language literature from 1 January 1980 to 14 November 2005. The 17 empirical studies identified were classified according to the level of incentive (for example, physician, group, payment system) and the type of quality measure rewarded. Thirteen of the 17 studies examined process of care measures, most related to preventive care. Five of the six studies of physician-level incentives, and seven of the nine studies of provider group incentives, found partial or positive effects on quality. One study found a negative effect on care for the sickest patients. Results in four studies suggested unintended side effects. No studies examined optimal duration of incentives or their sustained impact after termination. Overall, the authors observed that there were few empirical studies of the effects of explicit financial incentives on quality.

Rosenthal and Frank (2006) reviewed the literature on paying for quality in healthcare, including brief reviews of the pay-for-performance literature in other fields as well. In 2003, the authors examined the peer-reviewed empirical literature using five databases: MEDLINE, EconLit, ABI Inform, PsychInfo and the Social Science Citation Index. Additional citations were found by examining the reference lists of articles. The review focused on studies that assessed quality-based payment schemes. Studies were excluded that assessed the impact of payment systems on quality of care. The authors located seven published, peer-reviewed empirical

studies of the effects of paying for quality in healthcare. Another study located by the review related to contracting for substance abuse treatment, but the rewards were not spelled out so it was excluded from the review. The authors concluded that the empirical foundations for pay-for-performance in healthcare are 'rather weak'. There were only two positive findings and studies with the strongest research designs were more likely to find no impact related to financial incentives. However, the studies were narrowly focused and tended to relate to preventive care. Their implications for more recent pay-for-performance initiatives are not clear.

Scott and Hall (1995) reviewed the effects of different payment methods on GPs using a variety of measures of costs and outcomes of care. Four sources were used in searching the literature: MEDLINE, Social Sciences Citations Index, citations in articles received and citations known to authors. Studies were identified that examined actual changes in GP reimbursements or differences in GP reimbursed in different ways. The authors did not summarise their findings across these studies. Their main conclusion was that, based on the literature, it was not possible to make recommendations about optimal payment systems. Much more research is needed. According to the authors the most 'fundamental criticism' of the literature was that it didn't say whether patients were better or worse as a result of reimbursement changes. Only one study attempted this, comparing actual practice with clinical guidelines. This study, by the Department of Health in the U.K. (1991), found that GPs were more likely to hit some payment targets when paid specifically to do so. However, this was a before–after study with no controls and unclear data sources.

Town *et al* (2005) reviewed the impact of financial incentives on preventive care delivery. A unique aspect of the review is that it is limited to randomised trials. There were eight different financial interventions identified in the review. The incentives included direct payments or bonuses to providers, as well as more diffuse incentives. The authors searched EconLit, Business Source Premier, PsychInfo and MEDLINE. Reference lists were reviewed to identify other articles. The search focused on English language articles published from 1966 to 2002 that addressed primary or secondary prevention or health promotion. Studies using interventions with multiple components, where it was not possible to identify the effect of financial incentives, were also excluded, as were studies that compared outcomes under different payment systems. Two independent reviewers abstracted each article. Only six studies met the inclusion criteria and they generated eight different findings. Of the eight different financial incentives reviewed, only one led to a significantly greater provision of services. The authors noted that this doesn't necessarily imply that financial incentives won't motivate physicians to provide more preventive care. The incentives in the study were weak as the rewards were small. They concluded that small rewards probably won't motivate doctors to change their practices with respect to preventive care.

SUMMARIES OF EXPERIMENTAL STUDIES OF PAY-FOR-PERFORMANCE

Fairbrother *et al* (1999) examined the effect of different financial incentives on immunisation coverage, specifically the percentage of children up-to-date on a variety of immunisations. Physicians were assigned to one of three groups: bonus and feedback, enhanced fee-for-service and feedback, and feedback only. Physicians were randomly assigned to the three groups and immunisations were measured at three points in time, approximately four months apart. Nine neighbourhoods in New York City with the highest poverty rates were selected as study sites. Eighty-three physicians were invited to join the study and 61 accepted. Data were collected through chart review. Logistic and linear regression models were used to evaluate outcomes. There was a 25 per cent improvement in up-to-date immunisations in five categories for the bonus group, with no significant changes in the other groups. Much of the improvement appeared to be the result of better documentation.

Fairbrother *et al* (2001) conducted a follow-up to a previous study to analyse whether bonus payments and enhanced fee-for-service improved immunisation rates for children, specifically the percentage of children with up-to-date coverage on immunisations. Bonus payments were $1000 and $2500 for 30 point and 45 point improvements, $5000 for reaching 80 per cent up-to-date coverage and $7500 for reaching 90 per cent up-to-date coverage. In the enhanced fee-for-service group, physicians received $5 for each vaccine administered within 30 days of its due date and $15 for each visit at which all due vaccines were administered. A control group received feedback. A previous study by the authors left questions unanswered including: would the improvements in a bonus group continue, would actual practices (as opposed to documentation) improve over time and would the enhanced fee-for-service group begin to have an impact? Also, the previous study focused on inner city children served under Medicaid. This study included all payers. Incentives were given to 57 randomly selected physicians in New York City four times at four-month intervals based on performance for 50 randomly selected children in their practices. Logistic regression models and linear regression models were used to analyse the data. The lower response rate in this study (compared with the authors' previous study) was a limitation. Both types of financial incentives increased documented immunisations. The authors concluded that the incentives were not sufficient to overcome entrenched physician behaviour patterns and that true immunisation coverage was higher than documented in charts.

Grady *et al* (1997) evaluated the success of three different approaches designed to increase referrals by primary care physicians of patients 50 years and older for mammograms. The experiment added what the authors call a token reward for referrals to a strategy of 'cue enhancement' and education. The reward was a cheque based on the percentage referred in a given audit period (for example, $50 for a 50 per cent referral rate). In order to have a

comparison group, the rewards for some were not initiated until the second year of the experiment. The study was based on a randomised trial involving 61 practices in Dayton, Ohio and Springfield, Massachusetts over a three-year period resulting in a sample of 11,426 patients. The actual years covered in the analysis are not mentioned. Various statistical techniques were employed, primarily repeated measures ANOVA (analysis of variance), to test for statistically significant differences among the groups. Chart stickers were effective in increasing referrals. Peer-performance and feedback, combined with a reward, did not increase referrals over cueing alone. The authors speculated that the reward offered may have been too small and isolated to have had an impact.

Hillman *et al* (1998) evaluated a randomised controlled trial of an intervention intended to improve compliance with four preventive care screening exams for women 50 years and older, with financial incentives for physicians being part of the intervention. The three intervention sites with the highest compliance scores received a full bonus of 20 per cent of capitation. The three with the next highest scores and the three that improved the most got a 10 per cent bonus. Bonuses ranged from $570 to $1260 a site with an average of $775 per audit. Seventeen of the 26 sites received a bonus. Half the 52 primary care sites received the intervention, which included written feedback along with the financial bonus. The study was conducted from 1993 to 1995 in Philadelphia. Tests for the significance of differences in group means were carried out. Financial incentives and feedback did not improve physician compliance. The magnitude of the incentive may have been too small or the physicians may not have been aware of the change in incentives. Both groups saw dramatic increases in preventive care during the study period reflecting national initiatives.

Hillman *et al* (1999) conducted a randomised trial of two different interventions, one of which involved a financial incentive, to improve paediatric preventive care in a Medicaid population. The three practice sites with the highest total compliance scores with recommended practices received the full bonus of 20 per cent of capitation. The next three received a 10 per cent bonus, as did the three sites showing the greatest improvement, provided scores increased by at least 10 per cent. Bonuses ranged from $772 to $4682, with an average of $2000. Thirteen of 19 sites received at least one bonus and six sites received two bonuses. The purpose of the study was to determine if a system of semi-annual assessment and feedback, coupled with financial incentives, could improve paediatric preventive care guidelines as evaluated by semi-annual chart audits from 1993 to 1995. Fifty-three primary care sites in Philadelphia were assigned to three groups: feedback plus incentive, feedback only and a control group. Chart audits were performed at six-month intervals for 18 months. A statistical comparison of means was carried out. Neither intervention resulted in improved care. The authors noted that only 56 per cent of sites reported awareness of the programme despite repeated mailings.

Kouides *et al* (1998) conducted an empirical evaluation of the impact of a 1990 Medicare influenza project set in Rochester (New York State), with randomisation of physicians to a control group and an incentive group. Physicians could receive an additional $0.80 per shot or $1.60 per shot if practice immunisation rates of 70 per cent and 85 per cent were achieved respectively. The study took place in 1990 and 1991. Multiple regression techniques were employed in the analysis of physician reports of immunisations. A survey of physician offices was conducted to gather data on practice characteristics. The mean immunisation rate for practices in the incentive group was 68.6 per cent compared with 62.7 per cent for the control group. The median improvement was 10.3 per cent in the incentive group and 3.5 per cent in the control group. The authors conclude that, although the financial incentive was modest, it improved immunisation rates by about 7 per cent.

SUMMARIES OF EVALUATIONS OF P4P INITIATIVES OF HEALTH PLANS AND OTHER PAYERS

Physicians

Amundson *et al* (2003) analysed a programme to reward physicians in a single HMO to advise smokers to quit. Physician groups received bonus payments for achieving target scores on various quality indicators, including providing advice to patients to quit smoking. The authors did not indicate the amounts received by groups specifically for improving in this area. Audits of 14,489 ambulatory patient records were undertaken in 19–20 medical groups from 1996 to 1999. Statistical tests of before–after group means were conducted. Identification of tobacco use in patient records increased from 49 to 73 per cent, and advice to quit increased from 32 to 53 per cent. The number of medical groups in which 80 per cent of patient targets were met increased from zero to eight. The impact of financial incentives by themselves on provider behaviour was difficult to determine because the strength of the incentive was not clear, the incentive at the individual physician-level (as opposed to group) was unclear and the intervention was multi-faceted.

Armour *et al* (2004) conducted a retrospective claims analysis of the impact of physician eligibility for receipt of a bonus payment and performance of colorectal screening. A year-end bonus program was implemented for physicians in a managed care plan. Physicians received bonus payments for conducting colorectal screening for patients who turned 50 years of age. The exact nature of the bonus, including the amount, was not described in the study. The key study variable was "eligible to receive a bonus" with not all managed care physicians qualifying to be eligible. The health plan treated these criteria as proprietary. A multinomial logistic regression model was used to estimate the impact of physician eligibility for the bonus on patient receipt of colorectal screening controlling for patient and physician practice characteristics. Screening use increased significantly in the year after the bonus program was introduced; a 3 percentage point increase, or a 12.8 per cent relative increase. There was no comparable increase for patients of physicians not eligible for the bonus program.

Ashworth *et al* (2005) conducted a multivariate analysis of the relationship among factors related to physician incomes and achievement of quality performance indicators in an inner city health authority in the U.K.. The income of GPs depended on the number of patients in the practice, staff expenses and payments for performance. The time period was two years before new, higher payments for performance were instituted in the NHS. Data were collected on 151 practices in an inner city health authority for 2001 and 2002. Regression analysis, including path analysis, was used to explore relationships. The authors concluded that GPs were able to maximise their incomes by taking on more patients. Achievement of performance targets had little impact. Higher staff budgets were associated with better performance on quality

indicators, suggesting that the rewards for performance, which were not large, were offset by the higher costs of achieving higher quality.

Beaulieu and Horrigan (2005) estimated the impact of a managed care organisation's programme, which combined financial incentives and practice support, on the quality of diabetes care. Physicians who met targets or demonstrated significant improvement received a bonus. The largest payment was equivalent to a 12 per cent increase in their per member per month payment (true for both fee-for-service and capitated physicians). Actual payments ranged from $3000 to $12,000. Data on patient outcomes were self-reported by physicians, with limited audits of medical charts. The control group data were collected as part of the health plan's HEDIS (Healthcare Effectiveness Data and Information Set) reporting. Analysis consisted of statistical comparison of group means before and after the programme. There were significant improvements on five of six process measures. Thirteen of 21 physicians received a financial award. Of the eight not receiving rewards, six improved their scores. There was no evidence that quality declined in areas of care not being measured. Self-selection of physicians into the pilot programme and the small sample size limits the ability to generalise from the results. It also is impossible to determine the marginal effect of the financial incentive because it was implemented along with other practice supports for diabetes care.

Campbell *et al* (2007) assessed trends in quality of care indicators in physician practices in England before and after introduction of a pay-for-performance programme for GPs in 2004. GPs received payments based on the number of points they garnered in the course of a year. Points were awarded for practice structures supporting quality, process of care measures and access measures. Data from 1998, 2003 and 2005 were collected for 42 primary care practices in England for clinical indicators associated with coronary heart disease, asthma and Type 2 diabetes. Trend analysis was conducted for indicators that were eligible for reward under the pay-for-performance programme and also for some indicators that were not. Several different statistical methods were used to test the robustness of the findings. There was a statistically significant, but modest, increase in the trend rate for asthma and Type 2 diabetes indicators after the introduction of the pay-for-performance programme. The lack of a significant increase for coronary heart disease could be due to the fact that scores on these indicators were high prior to the pay-for-performance programme. There was no difference in the trend rates for indicators subject to pay-for-performance and for those that were not. The authors suggest that their analysis may underestimate the impact of pay-for-performance as practices may have implemented some changes in 2003 in anticipation of programme implementation. The lack of a difference between the trends for the two groups of indicators suggests that increases may not be due to pay-for-performance. Alternatively, practitioner attempts to improve scores on pay-for-performance indicators could have had a beneficial 'spillover' effect on other non-measured components of quality. The authors conclude that their results support the view that

pay-for-performance can 'make a useful contribution to improving quality' as 'part of a comprehensive quality improvement program' (p 189).

Campbell, McDonald, and Lester (2008) used interviews to explore physician and nurse beliefs and concerns subsequent to implementation of the U.K.'s pay-for-performance program. The pay-for-performance program in the U.K. rewards general practitioner practices for accumulating points by achieving target levels of performance relating to clinical quality, practice administration and other areas. Practice bonuses are directly related to the number of points practices accumulate. Forty-three health professionals (22 physicians and 21 nurses) in 42 practices were invited to participate in the study. The 42 practices employed 110 physicians and 71 nurses. Interviews were conducted between February and August 2007. The authors found agreement that the incentives had been enough to motivate changes in behavior. There was some resentment on the part of nurses that physicians tended to keep the bonus payments rather than distributing them to practice employees. There was also concern that trying to achieve points had changed the nature of the physician visit, possibly making it more difficult to address the patient's full agenda. There was some indication of reduced continuity of care and greater care fragmentation.

Casale *et al* (2007) analyzed a financial incentive program designed to reward increased quality within a single managed care plan. A fixed price for care of coronary artery bypass surgery was implemented, combined with target performance standards that were part of the "care package". The authors sought to determine if the new payment per episode of care approach, combined with the introduction of prompts in an electronic records system and a patient engagement program, could improve quality of care. Mean values were compared before and after program implementation for statistically significant differences. Care improved from an initial 59 per cent of patients receiving all 40 best practices to 86 per cent after 6 months. There were improvements in 30 day clinical outcomes but only "likelihood of discharge to home" was statistically significant.

Chiang *et al* (2002) described changes in reporting of TB in Taiwan from 1995 to 1999. Clinicians and hospitals received NT$250 for each confirmed case of TB reported. The authors plotted the number of reported cases from 1995 to 1999. The payment for reporting began in 1997. Changes between various reporting periods (six months) were calculated. There were no tests of significance reported. The incentive programme appeared to have its intended effect. There was a 47 per cent increase in reporting the year that the programme was instituted. However, respectively in 1998 and 1999 the number of reported cases declined slightly (7 per cent and 3 per cent). The study found an impact that was attributed by the authors to incentives. The result was probably significant, but no tests were performed, nor were any data presented concerning the nature of the increased number of reports.

Christensen *et al* (2000) carried out an evaluation of an intervention among pharmacists in the State of Washington that involved a financial incentive for providing cognitive services to Medicaid recipients at the time prescriptions were filled. Compensation was $4 for interventions up to six minutes and $6 for longer consultations. All pharmacies also received $40 per month for documenting the cognitive services they provided. Pharmacies were randomly assigned to a study (110) or control (90) group. Cognitive services documentation was audited for completeness and consistency. There was a significant difference in the number of cognitive services per 100 prescriptions (1.59 versus 0.67) and 75 per cent of consultations were less than six minutes. The authors do not state expectations directly, but imply that more consultative services results in better quality care, especially in improving patient safety.

Chung *et al* (2003) conducted a qualitative and quantitative assessment of a physician recognition programme employed in the Hawaii Medical Service Association. Physicians received points for achievement relative to quality indicators, patient satisfaction, business operations and utilisation of services. Physicians were ranked and the average incentive reward ranged from 0 to 5.5 per cent, with an average of 3.5 per cent. In 2001 a bonus of up to $3000 was added for practitioners who improved scores. There were payment caps to avoid higher payments to high-fee specialists. Administrative claims data were used to measure the quality indicators and utilisation. A survey was used to collect data on patient satisfaction. Non-parametric tests of statistical significance were conducted for the years 1998–2001. The programme started in 1997. The authors reported results on a subset of indicators (n=3). These are common measures but there is no explanation for why they were chosen for use in this particular case. There was improvement in ACE inhibitor use and in haemoglobin A1c testing. The results were mixed regarding improvement for an immunisation measure, a finding the authors attribute to external factors. There was no control group, so it is not clear if the improvements were due to the compensation programme. Only a subset of results is provided.

Collier (2007) compared performance by a hospitalist group before and after a reward system for quality was instituted, with hospital groups not under the same contractual incentives used as comparison groups. The exact nature of the reward system was not described. The author notes that rewards were tied to a variety of different types of performance measures, including access, timely records completion, meeting attendance, and meeting quality standards. Performance levels were tracked before and after the contract was put in place and compared with the performance of 2 groups not subject to financial incentives. After one year the hospitalist group under contract improved in all administrative areas, with no similar improvement in the comparison groups. All groups improved by a similar amount with respect to clinical indicators. Because improvement in these areas requires changes in clinical processes, the author speculates that it may take longer than one year to occur.

Curtin *et al* (2006) analysed the cost savings from a pay-for-performance programme directed at physicians providing diabetes care. Payments from a health plan to an individual practice association (IPA) withheld dollars which were then returned to the IPA if it met target performance levels. Each year about $15m of these withheld funds were distributed to 3700 participating physicians, specialists as well as generalists. An average primary care physician's distribution ranged from $6000 to $18,000 annually across all performance measures. Diabetes care was one component of the overall performance score on which payout was based. Historical trend data (2000–02) were used to estimate what the costs of care would have been for diabetes patients in 2003/04 in the absence of the pay-for-performance programme, and this was compared with the cost of the diabetes programme. Claims data provided by the health plan were used in the analysis. Savings were calculated from the perspective of the health plan. The authors found a positive return on investment of 1.6 to 1 in 2003 and 2.5 to 1 in 2004. The most significant cost reductions occurred in the area of hospital care. The authors pointed out that in most instances the pay-for-performance programme essentially rewarded physicians for providing more care for their patients with diabetes, presumably adding to direct treatment costs. Thus, the positive rate of return was more impressive than if achieving the performance goals had required no additional treatment or reductions in treatment.

Damberg *et al* (2005) presented early descriptive findings of the Integrated Health Association's pay-for-performance initiative, with discussion of design and implementation issues. Health plans used a common set of measures drawn from HEDIS to reward physician groups for performance, with public report cards distributed at the same time. Improvements in measures were reported. Tests of significance were referred to but specific results were not provided. Data were from the first reporting year (2003). There was significant improvement on at least four clinical measures for three quarters of the reporting groups. This article focused more on design and implementation issues than on analysis of improvements in quality measures.

Doran *et al* (2006) examined the first year experience of family practice doctors in the U.K. in achieving targets under the NHS's new pay-for-performance scheme. In 2004 the NHS committed about $3.2b in new funding for three years for a pay-for-performance programme for family practice doctors. Physicians were rewarded for their performance on 146 quality indicators relating to clinical care for ten chronic diseases, organisation of care and patient experience. Points were awarded on a sliding scale within a payment range, with payment limited to $133 per point awarded in 2004/05, adjusted for disease prevalence. In that period, the maximum that a GP could receive from the programme was $139,400. Data were extracted from a national computer database. Data for exception reports was imputed. Linear multiple least-squares regressions with robust estimates of error variance were used to estimate relationships. Fixed effects for practice location were used. The median practice achieved 95.5

per cent of the points available, in comparison to an expected 75 per cent. Achievement was higher in practices with a high ratio of family practitioners to patients, but all significant effects were small and only 20 per cent of the variance was explained by the regression models. The factor with the greatest effect was exception reporting. Physicians who excluded a large proportion of patients from the calculations performed better. The programme increased the gross income of physicians by an average of $40,200. There were no baseline data in the U.K. to use in the analysis, but there was evidence that quality was improving prior to the programme.

Doran *et al* (2008) used multiple linear regression analysis to examine factors that explain the rate of "exception reporting" in the U.K. pay-for-performance program. The pay-for-performance program in the U.K. rewards primary care physicians based on number of points attained in a given year, with the potential for the program to account for 25 per cent of a physician's annual income. In calculating points, physicians are allowed to exclude certain patients meeting predetermined criteria. This practice is called exception reporting, and there has been concern that physicians might use exception reporting inappropriately to generate higher payments. Average exception reporting was found to be much less that previous studies had suggested. This study was based on 2005-2006 data, while earlier studies were based on data from the first year of the program. Physicians excluded a median of 5.3 per cent of their patients. There was little association between patient and practice characteristics and exception reporting. While the average rate of exception reporting was similar to the first program year, the maximum estimated rate was substantially smaller. In all, exception reporting accounted for about 1.5 per cent of overall program costs. The authors concluded that fears of abuses of the exception reporting process seem founded.

Doran *et al* (2008) estimated the relationship between "deprivation" of census area in which a physician practices in the U.K. and change in clinical measures of quality under the U.K. pay-for-performance program. The pay-for-performance program in the U.K. is directed at primary care physicians. Physicians are awarded points for achievement of targets related to clinical and administrative clinical performance, and receive an addition to practice income depending on points received. Achievement was tracked over a three-year period on 48 clinical quality indicators for practices located in different census areas grouped by level of deprivation. Logistic regression was used to calculate the odds of practices being in the highest and lowest quintiles with respect to achievement. Multiple linear regression analysis was used to investigate associations between practice level characteristics and practice achievement. Median reported achievement levels grew in all practices in the first 3 years of the pay-for-performance program. In year 1, deprivation was associated with lower levels of achievement, but the gap between achievement and deprivation narrowed between the first and third years from 4.0 per cent to 0.8 per cent.

Doran and Fullwood (2007) described performance on measures relating to hypertension in the U.K. pay-for-performance program and discussed new evidence regarding exception reporting in the U.K. program. The pay-for-performance program in the U.K. contains several measures relating to hypertension, including blood pressure targets for patients with hypertension, coronary heart disease, diabetes mellitus, and stroke, along with blood pressure targets for all patients 45 and older. In the program, 173 points (16.5 per cent of the total) relate to blood pressure. A practice achieving all these points would receive a reward of $43,250, or an average of $13,000 per physician. The authors examine the percentages of practices achieving targets and changes in performance from year 1 to year 2. Also, in a secondary analysis, the degree of exception reporting was addressed. Rates of achievement were generally high for blood pressure indicators in year one, with a higher rate of achievement in year two than year one. The poorest performing practices in year one showed the greatest rate of improvement. Exception reporting rates were generally low and there was little evidence of widespread gaming of the reporting system. The authors noted that the targets for blood pressure were "less demanding" than would typically be found in blood pressure guidelines.

Ettner *et al* (2006) estimated the association between reimbursement incentives in 10 managed care plans and process measures for quality of care in diabetes treatment. The incentives faced by physicians were measured by proportion of compensation received from salary, capitation, fee-for-service and performance-based payment. A variety of performance measures were used, including receipt of dilated eye exams, foot exams, influenza immunisations, advice to take aspirin, and assessments of glycaemic control, proteinuria and lipid profile. Data were gathered in 2000 and 2001 through patient, provider groups, and health plan surveys and medical records reviews for 6,194 patients with diabetes. The analyses employed multi-level logistic regression techniques with random intercepts for provider groups and health plans. The most significant analytic problem related to high correlation between the payment variables and organisational type. When organisational type was not controlled for in the analysis, care processes were better when physicians were paid on a salary basis, and when quality/satisfaction scores were used to determine a portion of physician payment. The results were confounded by organisational type. Nevertheless, the authors concluded that, regardless of causality, use of quality/satisfaction scores to determine physician compensation 'may indicate delivery of high quality care for diabetes' (p 1222).

Felt-Lisk *et al* (2007) studied a Medicaid pay-for-performance demonstration involving contracting health plans in California. Providers were rewarded for achieving benchmarks for well-baby visits in the Medicaid population. Four of five plans offering new incentives offered bonuses to contracting entities based on the number of children who met well-baby visit guidelines. The fifth made payments directly to physicians using an existing bonus pool. A

difference in difference analysis was used to evaluate impacts where data permitted. Qualitative analysis was used to contrast the approaches taken by the Medicaid plans and the difficulties they encountered. Data covered the period from 2002 to 2005, with the payment years being 2003 to 2005. There were favourable trends overall in the number of well-baby visits, however the experience of the five plans in the study varied. The more successful programmes had greater rewards for providers and had better communication with providers about the incentive programme. There was little information provided in the article relating to the methodology used to estimate quantitative programme impacts.

Gene-Badia *et al* (2007) assessed whether implementation of an incentive scheme to improve quality and aid professional development had an impact on quality of professional life and patient satisfaction. Survey data were collected from 257 primary care teams and their patients in Catalonia, Spain in 2002 and 2003. Multivariate regression techniques were used to analyse the impact of financial incentives on 34 measures of quality of professional life and patient satisfaction with care and care facilities. Perception of support from management increased but so did perception of demands on health professionals. There was little evidence of an effect on patient satisfaction.

Greene *et al* (2004) evaluated a health plan programme to increase physician adherence to treatment guidelines for acute sinusitis in an IPA in Rochester (New York State). A scoring system was developed based on 20 per cent patient satisfaction, 40 per cent efficiency and 40 per cent quality measures. From 1999 to 2001 the percentage withhold was 15 per cent. In 2000, the withhold was reduced to 10 per cent for the top 5 per cent of performers and increased to 20 per cent for the bottom 5 per cent. Episodes of care were identified for acute sinusitis among 420,000 HMO patients between 1999 and 2001. Statistical process control charts were used to analyse changes over time, with statistical tests of the magnitude of the observed changes. The 'exception rate' decreased by 20 per cent, with most of the change being a decreased use of ineffective antibiotics. Given the multiple interventions involved, it was not possible to determine the contribution of financial incentives to the change.

Gulliford *et al* (2007) described trends in diabetes care in 26 South London practices prior to when the U.K.'s pay for performance program was implemented and in the first year after program implementation, supplemented by a cross-sectional analysis of factors that explained variation in practice performance scores during the first year of the program. Eighteen of the 76 clinical indicators in the program pertain to diabetes. They tracked trends in target achievement, without statistical testing, for the 26 practices. Regression analysis was used to assess factors associated with variation in practice performance on measures of diabetes care for all practices in the U.K. program. Among the 26 practices, there was improvement year to year in blood sugar and cholesterol control, with the largest improvements occurring in the

year that the pay-for-performance program was implemented. Practices in deprived areas were less successful in achieving targets in the first year of the pay-for-performance program.

Khunti *et al* (2007) compared the findings of a review of quality of care in diabetes treatment before implementation of the U.K.'s pay-for-performance program to outcome measures for diabetes care in the first year of the pay-for-performance program. Diabetes care accounts for 99 of 1,050 possible points under the pay-for-performance scheme. Maximum attainment would constitute less than 2 per cent of overall practice income. Their systematic review of the literature identified six studies of diabetes care before implementation of the pay-for-performance program that met inclusion criteria. The findings of these studies were displayed in tabular form alongside findings from the first year of the pay-for-performance program. The authors concluded that the incentives in the pay-for-performance program led to substantial improvements in many process of care indicators and all intermediate outcome measures. However, there was no control group and the study design did not permit statistical analysis of trends.

Kraft *et al* (2008) analyzed physician survey data to determine if the opportunity to gain insurance payments affected the probability that physicians followed a recommended treatment protocol for patients with diabetes (self-reported). Accredited facilities can receive a case-based payment for providing a recommended package of services to patients with tuberculosis. The probability of adopting the treatment protocol for diabetes was estimated using a binary logit model and maximum likelihood techniques. Training was found to be a better predictor of adherence to a treatment protocol for TB when the protocol was a significant departure from past practices, but financial incentives were found to be more effective for practices that have demonstrated clinical competence.

Langham, Gillam and Thorogood (1995) examined changes in the distribution of health promotion financial incentive payments after a programme was implemented in the U.K. in 1993 that focused payments on cardiovascular disease. Payments were associated with the performance of screening and the recording of risk factors. Previously, GPs were paid for holding health promotion clinics. The study examined the distribution of health promotion payments between health services authorities and between general practices. The retrospective study of payments included the periods before and after the change in payment approach. Payments were analysed for 78 practices in one authority and 85 in another. Changes in payments were calculated for two measures of relative need. Statistical comparisons of means were conducted. Health promotion payments were found to be more evenly distributed after the change. Practices in areas with the highest need lost more. In general, the resulting distribution was unrelated to need or treatment given after the change.

Larsen, Cannon and Towner (2003) assessed the impact of a disease management process developed by an integrated health system (Intermountain Healthcare – IHC) on diabetes care. The financial incentive was part of a broad-based care improvement effort that included many components. The incentive was not described in detail, but appeared to be relatively small, representing 0.5 to 1 per cent of physician compensation. About half of that incentive related to diabetes care. The authors reported improvement on the key performance measures that were significant and clinically important. Because of the broad-based nature of the programme, it was not possible to determine the impact of the financial incentives by themselves. However, the incentive was small, and therefore it does not seem likely that it was a major influence on behaviour.

Levin-Scherz, DeVita and Timbie (2006) analysed pay-for-performance contracts with physicians for diabetes and asthma care in Massachusetts. The incentive in the programme was applied at the network level. There was a withhold in provider contracts, often at 10 per cent of fees. In some cases there was the opportunity for bonus payments beyond the fee schedule. Withholds were returned or bonuses earned depending on performance relative to agreed targets. There was a variety of metrics, but the article focused only on performance on diabetes and asthma care. The analysis used claims data taken from the network's multi-year data warehouse. A difference-in-difference analysis was used for the state comparisons. The years included in the study were 2001 to 2003. There was improvement in the network's diabetes measures relative to the index plan in the state and relative to national plans. There was also improvement in asthma measures, but performance in this area started at a relatively high level. The authors noted problems in using claims data to reward performance and the network was looking forward to having access to patient electronic medical records.

McDonald *et al* (2007) conducted an ethnographic study to assess the impact of the NHS pay-for-performance programme on practice organisation, clinical autonomy and internal motivation of GPs and nurses. Data collection took place in two practices in deprived parts of north-west England. These practices had reputations for high quality care and high scores in the first year of the pay-for-performance programme. Observation was combined with interviews, informal conversations and document review. The authors concluded that implementation of financial incentives did not damage internal motivation of GPs, although nurses expressed more concern. Most GPs did not question the quality targets or their implications for clinical quality.

McDonald, Harrison, and Checkland (2008) conducted case studies of two practices in England following implementation of pay-for-performance. The pay-for-performance program in the U.K. is directed at general practitioners, who can earn points based on achievement relative to an extensive number of performance measures. The bonus payments that practices receive depend on the number of points they accumulate. The authors assessed attitudes and

behaviours in general practice, with an emphasis on mechanisms and perceptions of control, using observation, interviews and analysis of documentation. Attitudes towards the pay-for-performance program were found to be generally positive in the two practices, but there was discontent in the practices that employed a stronger surveillance system. Nurses who were given responsibility for achieving targets felt greater pressure.

Mehrotra *et al* (2007) conducted a multivariate analysis of survey data collected from physician practices in Massachusetts. Incentives were provided by health plans in a variety of areas including process of care and utilization. The surveys did not ask about specific program incentives. The multivariate analysis addressed the quality improvement activities undertaken by the physician groups. Survey data were collected through telephone interviews with 79 leaders of physician practices in Massachusetts. The leaders were asked about the types of pay-for-performance incentives they faced and their quality improvement activities. A descriptive analysis of the survey data was provided, as well as a multivariate analysis of the association between incentives and the practices' quality improvement activities. Eighty-nine per cent of practices reported incentives for at least one commercial health plan, typically tied to HEDIS reporting measures. About 2/3 of reported incentives were related to utilization measures. Pay-for-improvement incentives tied to HEDIS measures were positively associated with group quality improvement initiatives.

Millet *et al* (2007) conducted a population-based longitudinal study of the effect of the U.K. pay-for-performance program on delivery of smoking cessation advice and on prevalence of smoking among diabetic patients. The U.K. pay-for-performance program rewarded general practitioners points for achieving clinical and administrative targets. The size of the bonuses paid to physician practices depended on the number of points achieved. The authors analyzed change in recording of smoking status, in documented smoking advice given by the practice, and in the prevalence of smoking among diabetic patients in a single primary care trust containing 36 primary care practices. Conditional logistic regression analysis was used to analyze the data and adjusted odds ratios were reported. More patients with diabetes had their smoking status recorded, and there was significant improvement in documented smoking cessation advice (48 per cent to 83.5 per cent). The prevalence of smoking decreased significantly from 20 per cent to 16 per cent.

Morrow, Gooding and Clark (1995) studied associations between a multi-faceted intervention and improvements in the preventive healthcare behaviours of physicians in an IPA. The only information provided by the authors regarding financial incentives was that a good score on preventive services increased reimbursement for physicians in the health plan. Chart audits of practices in a four state area were conducted from 1987 to 1990 (the number of practices was not provided). Confidence intervals were calculated. There were improvements in virtually all of

the preventive measures. The authors observed that they could not necessarily attribute the improvements to the plan's programmes, including financial incentives, as there were confounding motivations for change in physician behaviours.

Parke (2007) analyzed claims data to assess overall health care costs in an employed group before and after introduction of a pay-for-performance program. Physicians received a higher fee-for-service payment if they treated patients as recommended by an electronic reminder related to evidence- based treatment processes. The amount of the increase was approximately 10 per cent. The authors assessed net changes in fixed and variable expenditures by the employer and employees before and after the program was implemented, comparing mean values over two years, with no statistical analyses. Total expenditures declined even though there were benefit design changes and increased per unit payments for physicians. The contribution of the financial incentives for providers was unclear because many other changes, including patient incentive program, were implemented simultaneously.

Ritchie (1992) tracked immunisations in a single region in Scotland before and after introduction of a new contract for primary care physicians in 1990. In this contract, 'item of service' payments were replaced by target payments to encourage GPs to increase rates of childhood immunisations. The details of the payment change were not described. In the study region, this change was accompanied by a records system that provided feedback to GPs regarding their immunisation performance. Immunisation rates for 95 practices encompassing 313 GPs were calculated for children aged two and five on the first day of each quarter for the seven quarters ending in March 1990 and subsequent three-month periods to September 1991. The analysis was retrospective and descriptive and used data drawn from the computer records maintained by the Grampian region in Scotland. A variety of statistical analyses were conducted using linear, non-linear and logistic regression methods. The practices achieving immunisation rates of at least 95 per cent increased from 31 to 81 per cent for primary immunisations. Achievement of 95 per cent rates for pre-school booster immunisation increased from 23 to 64 per cent. The authors noted evidence of 'sustained improvement' but no change in overall trends. They suggested that the reasons for the change were complex and should not necessarily be attributed to the new contract and the change in financial incentives introduced by it.

Rosenthal *et al* (2005) evaluated the impact of a physician pay-for-performance programme implemented by a health plan. Beginning in July 2003, participants received a quarterly bonus of $0.23 per member per month for each performance target met or exceeded. The overall potential for a group with 10,000 health plan patients was $270,000 per year. This represented about 5 per cent of professional capitation paid by the plan and about 0.8 per cent of the group's overall revenue. The evaluation focused on three process measures of clinical quality:

cervical cancer screening, mammography and haemoglobin testing. Within the plan, some medical groups received pay-for-performance payments, while groups in another region did not. Data on performance were available before and after the programme was implemented. Generalised least squares techniques were used to estimate a difference-in-difference model. Compared with the groups not receiving a pay-for-performance payment, the groups receiving payment demonstrated greater improvement only in cervical cancer screening. Because payment was made for achieving benchmarks, groups that improved the least, because they started out at a high level, received the most bonus money.

St. Jacques, Patel, and Higgins (2004) assessed the impact of implementing a programme of physician profiling, reporting and incentives on the behaviour of anaesthesiologists. For each study month physicians were eligible to receive a variable financial payment of $0–500 depending on individual scores relative to one another. The payment was credited to the physician's personal expense account. Performance was tracked in five areas: percentage of first cases of the day at the room before or at start time, percentage of cases where preparation time was less than a target, percentage of cases delayed while waiting for anaesthesiology evaluation, percentage of cases delayed during anaesthesiologist controlled time and percentage of cases delayed while waiting for anaesthesiology attending. Thirty-one anaesthesiologists in a university hospital were tracked for six months. A statistical comparison of means was carried out. Compared to the first month, performance improved on most measures. Because the programme combined profiling with incentives it was not possible to determine the effect of incentives only. The authors did not relate their findings to patient outcomes.

Simpson *et al* (2006) analysed of the impact of a new payment scheme for GPs on recording of quality indicators for patients with stroke. The new payment system, introduced in Scotland in 2004, provided payments to practices that developed an accurate register of stroke patients and for the recording of smoking habits, blood pressure and cholesterol levels. There were also payments for reaching blood pressure control targets and other outcomes. Retrospective data from 310 (self-selected) of Scotland's 850 practices were obtained from a central database in 2005, including data for one year before the new incentive system was introduced and one year after. Binary logistic regression was used to calculate odds ratios for recording of data. Documentation increased from 32.3 to 52.1 per cent. There was a large increase among the oldest patients and most affluent patients. Women had larger increases in documentation than men. The authors noted that inequitable recording still persists, with lower recording for women, older patients and more deprived patients.

Srirangalingam *et al* (2006) conducted an empirical analysis of how referral patterns for diabetes care changed after introduction of the new financial reward system in the U.K.. Under

the new general medical services contract for primary care in the U.K., primary care physicians receive financial rewards for performance on diabetes-related quality indicators. Referrals from primary care to a hospital-based diabetes service before and after implementation of the new incentive system were tracked. The study setting was a deprived area of London. Referrals were tracked from November 2003 to November 2004. Statistical tests of significant differences at the 0.05 level were carried out. There was no significant impact on the total number of referrals to the specialty clinic, but there was a significant increase in referrals for poor glycaemic control. The authors concluded that the new contract led to an increase in referrals for patients with unacceptable glycaemic control along with a lower threshold for referrals.

Sutton and McLean (2006) assessed factors related to quality scores under a new U.K. primary medical care contract that pays GPs in part based on quality measures using a relatively complicated formula that the authors do not describe. Data were analysed for 60 practices in two NHS areas in Scotland serving a population of 367,000. Linear regression analysis was used to relate quality scores to various characteristics of the population, GP and GP's practice. The most relevant finding is that practices with higher incomes from other sources had lower quality scores. The authors speculate that the incentive effect of the new contract is weaker when income from other sources makes up a larger portion of practice income.

Whalley, Gravelle, and Sibbald (2008) conducted a statistical analysis of a longitudinal survey of physicians in the U.K. to assess changes in attitudes towards a pay-for-performance program. The pay-for-performance program in the U.K. is aimed at primary care physicians, who can receive bonus payments based on achievement of a wide variety of administrative and clinical targets. The study considered changes in physician responses to a survey that asked about job satisfaction, hours worked, opinion of the impact of pay-for-performance on quality, and other items. Ordinary least squares, fixed-effect, panel data, multiple regression models were used to analyze the physician survey data. The authors found improvements in satisfaction with work hours and remuneration. Physicians were more positive about the impact of the contract on quality of care than they had expected to be.

Young *et al* (2007) evaluated the impact of a financial incentive program for primary care physicians on five diabetes performance measures. Each physician had about 5 per cent of fees withheld and transferred to a performance pool. The money was distributed to physicians based on their performance relative to indicators of clinical quality, patient satisfaction, and practice efficiency. The possible return to an internist in 2003 was $5,500-$16,500. Two-way repeated-measures analysis of variance was applied for each performance measure, testing for statistically significant changes in performance levels and trends over three time periods. Comparisons were made with general trends in the state and nationally. The authors concluded that the overall improvements in performance reflected secular trends. A "modest"

one time improvement in physician adherence to eye examination recommendations was attributed to the program.

Institutions

Bhattacharyya, Mehta, and Freiberg (2008) conducted a multivariate analysis of hospital characteristics that predict hospital performance in the top 20 per cent of a pay-for-performance program related to hip and knee replacements. The CMS/Premier Hospital Quality Incentive Demonstration awarded hospitals performing in the top 10 per cent nationally with a 2 per cent addition to their DRG based payment. Hospitals in the top 20 per cent, but not the top 10 per cent, received a 1 per cent bonus. Hospitals were graded on three process measures and 3 outcome measures. The dependent variable in the analysis was whether or not a hospital was in the top 20 per cent. Variables for which univariate tests of association were not significant were dropped from further analysis. A logistic regression equation was estimated that incorporated the remaining variables. Hospitals in the top 20 per cent in the pay for performance calculations were more likely to be located in the Midwest and be teaching hospitals. Neither hospital size nor revenues were associated with top performance. Volume of surgeries was associated with being in the top 20 per cent.

Type of Study: Observational

Cameron, Kennedy, and McNeil (1999) analysed the impact of a programme of bonus payments for 21 hospitals for improved provision of emergency services. Beginning in 1995, 21 public emergency departments in Victoria, Australia were given bonus payments at the beginning of each fiscal year. They were required to return portions of the bonus if they were unable to meet targets for emergency care. The payments started at AUD$7.2m in total, and increased to AUD$17m by 1997/98. The targets related to areas of performance such as ambulance bypass, waiting time for patients with different levels of emergency and access block (patients waiting more than 12 hours for admission to a hospital). The authors used regression analysis to examine performance on the set of payment measures for two years before and three years after the bonus programme was initiated. The data were self-reported by the study hospitals and not audited. There was no explanation regarding how the authors specified the regression equations and carried out their statistical tests. The authors found that performance improved in all areas. All the results were significant except for the reduction in access block. These results were sustained over the three-year post-intervention period. The authors attributed the success of the incentive programme in part to the fact that it was developed collaboratively with local providers of emergency care.

Glickman *et al* (2007) analysed whether a hospital pay-for-performance programme implemented by Medicare improved care for patients with AMI. Hospitals in the two highest deciles of performance received a reimbursement bonus while those in the lowest decile risked future financial penalties under Medicare's Hospital Quality Incentive Demonstration, which began in 2003. In the first two years, payments totalling $17.55 were made across five clinical conditions, one of which was AMI. In the first year, 123 hospitals received payments; 115 received them in the second year. Data were used for 500 hospitals already participating in a quality improvement initiative (CRUSADE); 54 of these were in the Medicare pay-for-performance initiative, allowing for the creation of a control group of 446 hospitals. Data covered a period before and after the pay-for-performance demonstration. Each hospital collected data and submitted it to a central database. Six different processes of care measures were evaluated as the primary outcome measures. The study also included eight measures of care that were not included in the measures Medicare rewarded as part of the demonstration. There were slightly higher rates of improvement for two of the six measures rewarded by Medicare: aspirin at discharge and smoking cessation counselling. There was no significant difference in a composite that included all six measures, nor was there any significant difference in a composite consisting of all measures not rewarded by Medicare. The hospitals in the analysis were all volunteers and were already committed to improving treatment for patients with AMI. The authors concluded that, while there was no evidence of improvement due to pay-for-performance, neither did they find any adverse effects.

Grossbart (2006) evaluated the impact of the CMS (Centers for Medicare and Medicaid Services) demonstration project on performance improvement in hospitals, using hospitals in a single multi-hospital system. Under a three-year demonstration programme instituted in 2003, 278 hospitals were given financial incentives based on 35 quality measures in five clinical areas. For each clinical area, hospitals with composite scores in the top 10 per cent received a 2 per cent bonus payment on top of normal payments. Hospitals in the second decile received a 1 per cent payment. There was a slight downside risk in the third year for hospitals that did not perform above threshold quality scores. The setting was Catholic Healthcare Partners, which has its headquarters in Ohio. Some of its hospitals participated in the pilot, while others did not; hence, these acted as a control group. Analysis was limited to three of the five clinical areas: AMI, heart failure and pneumonia. Performance in the first year (2004) was compared with the previous year using composite scores. The study was based on care provided to 28,925 patients. Data were obtained from the database of the hospital system. A comparison of mean values was conducted. The pace of quality improvement in the pilot hospitals was found to be slightly greater than in the control group.

Karve *et al* (2008) conducted a statistical analysis of the relationship between a hospital's performance in Medicare's P4P program and the proportion of patients who were African American. Medicare provides financial incentives to hospitals whose care performance ranks in the top 20 per cent, and in the top 10 per cent, in specific disease categories. Hospital performance on measures of acute myocardial infarction, community-acquired pneumonia, and heart failure was analyzed for the second quarter of 2004 and the first quarter of 2005. Multivariate logistic regression was used to determine the association between percentage of African American patients in a hospital and the likelihood that the hospital was in the lowest or highest quintile of performance. There was an inverse association found between per cent of African American patients and performance related to acute myocardial infarction and community acquired pneumonia. The authors concluded that the P4P program may be exacerbating existing racial ethnic disparities in hospital care.

Lindenauer *et al* (2007) assessed the impact of a Medicare hospital pay for performance initiative on four composite measures of quality of care. Hospitals performing in the top decile on 33 quality measures relating to five conditions received a 2 per cent bonus payment. Those in the second decile received a 1 per cent bonus and hospitals not performing above the level of hospitals in the lowest two deciles (established in the first year) were penalised from 1 to 2 per cent. Bonuses averaged $71,960 per year. The set of hospitals in the study included 613 hospitals that voluntarily reported information about quality of care through a national public-reporting initiative; 207 of these also were participants in the Medicare pay-for-performance demonstration. Changes in performance were compared for the two groups of hospitals, using multivariate methods to control for differences in hospital characteristics. After adjustment for hospital characteristics and baseline performance, pay-for-performance was associated with improvements from 2.6 to 4.1 per cent over two years. The main share of bonus payments went to hospitals with the highest performance at baseline, but hospitals at all levels of baseline performance improved. The authors view the improvements as modest and acknowledge that the hospitals volunteered and that their attempt to control for 'volunteer bias' may not have been entirely successful. In analyses that did attempt to control for this possible bias, effects were smaller.

Nahra *et al* (2006) estimated the QALYs gained in a patient population hospitalised for heart treatment, relative to the money spent by a health insurer in pay-for-performance payments to hospitals. A variety of assumptions needed to be made to generate the estimates in the paper. Eighty-five hospitals in Michigan received about $22 million in a four-year period. The hospitals were paid for achieving minimum levels of compliance with accepted clinical standards for two heart conditions. The incentive payments were add-ons to the hospital's DRG-related payments from BCBS of Michigan (a national health insurer). The maximum add-on for heart care was 1 to 2 per cent between 2000 and 2003. Thresholds for receipt of payment were increased from

year to year to encourage continuous improvement. The authors translated the measured improvements into estimated years of life gained relative to cost of the programme. Data on costs were collected from BCBS. Process measures were collected over a four-year period from 2000 to 2003 based on hospital self-reports. The authors used estimates from the literature to convert the process improvements into estimates of QALYs gained. The cost per QALY was estimated to be between about $13,000 and $30,000, which the authors observed is well under the consensus measure for value of a QALY, indicating that the initiative was cost effective.

Nalli *et al* (2007) conducted a descriptive study of a hospital pay-for-performance programme implemented in Maine in 2005. Hospitals received payments from a fund established by the hospitals for reaching an agreed performance level and then bonus payments from employers based on performance against 22 measures encompassing patient satisfaction, patient safety, clinical effectiveness and efficiency. Qualitative data were collected from programme participants and data on distribution of funds were collected from secondary sources. Six of the ten participating hospitals received payments totalling $89,645. The participants believed that the programme led to care improvements in their hospitals. The programme was not continued, but it was expected that health plans would use the measures developed by the programme in their pay-for-performance efforts.