Title: Pay for Performance: An International Assessment

By

Pauline Vaillancourt Rosenau, Ph.D.
Professor
Management, Policy, and Community Health Division
University of Texas, Houston - School of Public Health
1200 Hermann Pressler Dr. RAS E 915
Houston 77030-3900 Texas USA
email: Pauline.Rosenau@uth.tmc.edu

and

Dr. Christiaan Lako
Public Administration
Radboud University Nijmegen
Nijmegen, The Netherlands

ABSTRACT

Facing increased health care costs in a context of global fiscal crisis some health policymakers are employing Pay for Performance approaches to reduce health system costs and improve quality. This paper traces its origins, addresses the underlying assumptions of this approach, summarizes its theoretical advantages and disadvantages, and reports on the evidence as to its successes and failures to date. A review of the literature is employed to assess differences and similarities in results with Pay for Performance across the industrialized countries. The conditions in which it appears fruitful and situations where it is ineffective are delineated. Special attention is given to unexpected and to unintended consequences of Pay for Performance policies. Both critical qualitative appraisals and quantitative studies are included in this assessment with recommendations to inform policy. Research in other fields predicts that zero-sum Pay for Performance may not be as successful in improving quality as is anti-anxiety, positive competition. Within this structural form all can “win” if they meet the standard and no one is punished for not meeting the standard.
Pay for Performance: An International Assessment

What is “P4P” and Why Definitions Matters

Pay-for-Performance is having an enormous impact in the US; it is the most important new development in the healthcare sector in the USA since managed care (HMOs) and capitation. It aims to change behavior of those working in hospitals, doctors (primary care physicians), nurses, home health care agencies, psychiatric care providers, and patients with rewards and/or punishments. Pay-for-Performance – also written as "P4P“-- is a system for rewarding those involved in the delivery of healthcare services (doctors, medical groups, and hospitals) who adhere to specified clinical guidelines, quality performance measures, and efficiency-related indicators (cost). Performance can be defined as increased volume of patients treated, reductions in waiting time, etc. It can refer to designated processes (entering patient information in electronic medical records) or to specific outcomes (lower cholesterol, for example).

A P4P reward or bonus might take the form of a salary add-on to the general fee-for-service system (DRG for hospitals) for doctors. It can be additional payments for hospitals that perform well under the terms of a given metric. Pay for Performance incentives can also mean lower pay or reduced compensation for poor results. It could lead to imposition of a penalty for failure to respect guidelines or to meet other specified criteria.

Definitions of P4P are important because they can influence results by setting the basis for judgments. P4P quality may be defined in line with the U.S. Institute of Medicine’s understanding of quality as, effective, patient-centered, timely, efficient, and equitable care. Other definitions might lead to different results, for example, reduction in errors. The results of
P4P and who gets rewarded might be different if the perspective is population health (community health) rather than clinical care of individual patients. If patient satisfaction is defined as a quality measure that is to be rewarded, then customer relations may become more important to providers than clinical outcomes. Providers will generally focus on whatever aspect of health care that is rewarded on the P4P metric.

Value assumptions underlie P4P reward criteria. Definitions and what is rewarded under P4P includes value assumptions. For example, if P4P looks to practice guidelines, then patient satisfaction is not as important as outcomes. If medicine is a business, then patient satisfaction counts a lot because within a business model a patient is assumed to return to the provider if the experience is positive. Of course the two are not unrelated, but P4P is selective.

**Historical Roots, Origins, and What Can Be Learned from the Past in Other Sectors**

P4P was employed in other fields of social endeavor long before it became important in health care. Behn points out that “performance measurement is not an end in itself.” (Behn, 2003, p 586). He suggests that different performance measures are appropriate for various managerial purposes. This important lesson is sometimes forgotten. Education is one notable area where there has been much experimentation with P4P. Teachers are paid more if the class improves. Experience with P4P in education has been mixed. Critics argue that it encourages a “Teach to the test; ignore the rest” mentality, where teachers only teach what is going to be on the test at the end of the year.

Despite these criticisms P4P is broadly supported across the political spectrum but few disagree that “you get what you target, whether you like it or not.” And if the criteria for rewarding teachers changes, behavioral change is thought to follow.
Who Favors Pay for Performance and Who Does Not?

Who supports Pay for Performance

As with most social policy, some advocate P4P and others oppose it. Supporters include a wide variety of stakeholders including private and public sector entities that actually pay for health services. Examples are private insurers and public payers such as Medicare in the US. Both have experimented with it directly or commissioned pilot studies. It has been argued that P4P is associated with the Republican party in the US and with the George W. Bush administration, but the Democratic administration of Barak Obama also supports it within the Medicare program (Kaiser Family Foundation, 2006)

Why do they support it?

The intuitive appeal of P4P is not surprising. It makes logical sense and appears to be a relatively simple solution to several of the problems facing the health sector in industrialized countries. There is an expectation that P4P will facilitate cost control (Wodchis, Ross, & Detsky, 2007). Business consultants advocate it because they believe it to be of help in cutting costs. And there is evidence that paying doctors to do "certain things" works in the sense that one will get more of what one is rewarding providers to do, if the reward is large enough (Greengarten & Hundert, 2006).

Who does not support P4P

Critics of P4P include a wide range of providers including physicians and hospitals that stand to lose money in the zero sum competition that is often a structural characteristic of P4P.
Some critics from the field of management contend that P4P is objectionable for philosophical reasons and because it has negative effects on the internal social dynamic of the organization. Others contend that the way P4P is structured is important in determining whether or not it is a positive innovation – for example win-win approaches to P4P are likely to have very different effects compared to zero-sum approaches. This is discussed below. Support or criticism of P4P may depend on how it is implemented.

**Criteria for P4P are Critical and There Is Several Different Views about It**

Deciding what criteria to reward in any P4P program is critical in determining both support for it and the results that accrue. This has proved to be one important area of controversy within the P4P innovations. Because P4P may change provider behavior and because it almost always sets in motion various tactics on the part of providers to game the reward system, it is critical that any P4P metric established should be valid and reliable.

P4P programs necessarily depend on simple, gross indicators (measures) for judging "performance" and they end up making doctors or hospitals focus on what they are being "tested" for and paid to do. One problem is that health care personnel may neglect other things, not part of the P4P metric, that are also important. For example, P4P quality ratings for hospitals sometimes give higher performance ratings if patients receive Beta Blockers after a heart attack. But there are many other things that would be happening and Pay for Performance must, by definition, use very simplified indicators.

Current knowledge of what is best practice might be used as a basis for rewarding providers, but this approach is more legitimate if guidelines are based on scientific evidence of appropriate care. However, there are few definitive bases for medical decision making and much
controversy surrounds guidelines (Sanghavi, 2008). In some cases what is "best" or "correct" is unknown. Practice guidelines are not perfect and there are many exceptions to each one.

New knowledge and additional research results mean that practice guidelines must be updated and the criteria for P4P consequently change. The above mentioned practice guideline that indicates the administration of beta-blockers for heart attack victims may be wrong and guidelines in the US were changed April 1, 2009 (Newman, 2009).

For example, best practice and guidelines changed for radical mastectomy for early stage breast cancer, hormone replacement therapy for all postmenopausal women, and the value of beta carotene vitamin supplements (Epstein, 2007). But none of these is considered the “standard of care” today. This means that what P4P rewards today, may not be the same and could be countermanded in the future. P4P becomes very complicated in these instances where “best practice” is evolving.

In some cases not all patients with the same medical problem should be treated in exactly the same way and this makes formulating P4P guidelines imperfect. Such distinctions are difficult to formulate and reward under most P4P systems. One way this is accomplished is to set up “exceptions” for regular P4P guidelines. But this introduces ambiguity and may undermine the intent of the guidelines especially where the physician believes that his or her approach, based on his or her personal experience, is superior to the guidelines that are the basis for a reward system.

Co-morbidities present another problem and complicate patient treatment protocols. It is almost impossible to formulate P4P standards that take into account that patients with many problems may have to be treated to optimize their condition, rather than according to the standard guidelines indicated for patients who have no other problems.
Who should set the basis for P4P is another area of controversy. Practitioner groups argue for a predominant role for their professional group in establishing best practice. However, insurers and payers, in general, find that practitioners do not take into account the cost of various approaches to treatment. They want more of a say.

**Does Pay for Performance Work?**

In the US the number of studies and pilot projects assessing P4P has multiplied in recent years. Taken together the results can be characterized as at best, mixed, and at worse, inconclusive. The following examples are drawn from P4P in the US to illustrate the divergence of research results on the topic of P4P. A 3 year study of 105,383 U.S. Medicare patients from 123 hospitals involved $17.6 million in bonuses for hospitals: Researchers report that “the pay-for-performance program was not associated with a significant incremental improvement in quality of care or outcomes for acute myocardial infarction.” (Glickman et al., 2007). The Centers for Medicare and Medicaid Service; undertook a 3 year demonstration project involving 266 hospitals and it awarded a total of 8.7 million to 115 hospitals. The researchers concluded that P4P “lifts quality.” (Abelson, 2007). The Centers for Medicare and Medicaid Service; experimented with 10 groups of physicians involving a total of 16.7 million as rewards for achieving P4P goals. The results were that “all 10 participating physician groups achieved target performance,” and apparently improved health outcomes. However, only a few groups reduced costs. This study was not peer reviewed and no control group was included (Centers for Medicare & Medicaid Services, 2008). Pearson and collaborators studied 5 health insurance plans in Massachusetts with different levels of payment incentives for P4P. The study concluded that “Overall, P4P contracts were not associated with greater improvement in quality compared
to a rising secular trend” (Pearson, Schneider, Kleinman, Coltin, & Singer, 2008). In 2006, The US Institute of Medicine published a synthesis of 17 studies about the effect of P4P on quality of health care. Of these, 5 showed improvement, 8 were mixed, 3 reported no effect, and one actually found that quality was reduced (Peterson, Woodard, Urech, Daw, & Sookanan, 2006).

**Explaining Discrepant Results: Why the Confusion on Results and What Are the Methodological Problems**

There are several explanations as to why results from P4P experiments in the US are so divergent. The samples involved are said to have been very small. Participation in these programs was not based on random samples of practitioners or hospitals – but rather voluntary participation. Change in performance under these P4P programs sometimes coincided with structural changes or institutional or organizational reorganization. The end result is a situation where observed change cannot be attributed to P4P alone and could be simply the result of the structural innovation (B. R. Golden, 2006; Pink, Brown, Studer, Reiter, & Leatt, 2006).

Another difficulty that could account for difference findings regarding the effectiveness of P4P is that improvement is observed in the study but every element in the population, even those not researchers with P4P, are improving on quality, cost, etc. over time. In this case researchers might conclude substantial performance improvement in the P4P experiment when in fact everyone was improving, even when P4P was not present. This situation underlines the importance of a control group but few P4P studies include one. In the absence of a control group what are called “secular trends” could overwhelm the effects of P4P.

Discrepant results across countries regarding P4P, according to some policy experts, is because doctors and hospitals in some countries were not paid enough. This would explain the greater failure rate of P4P in the US compare to other countries such as Britain where rewrqrd
were far greater (Davis, 2008; Doran, Fullwood, & Gravelle, 2006). In the US, P4P associated rewards are seldom more than 5% of a physician’s salary (Advisory Board Company & Foundation, 2008). The idea is that if you pay people **enough**, eventually it will have an impact on their behavior.

**Unintended Consequences of Pay for Performance**

Enough experimentation with P4P in the health sector is now available to identify some unanticipated effects of these programs – others may arise and can be anticipated thought they have not been documented. All reimbursement systems encourage gaming of one sort or another and P4P is no exception. How this works will vary from P4P plan to another and it depends on how the P4P is structured. Gaming has been found to be more of a problem in the US than in the UK where P4P has been said to encourage cheating and fraud, exaggerated forms of gaming (Galvin, Delbanco, Milstein, & Belden, 2005; Hamblin, 2008; Hayward & Kent, 2008). There is even a guide on how to do game the P4P system, published in one of the most highly respected medical journals in the US (Hayward & Kent, 2008)!

P4P has been found to reduce cooperation within the workforce if structured in certain ways (discussed below). In other cases it has been met with indifference and reduced morale, especially by those that are on the low performing end of the continuum. This may because these individuals or organizations become discouraged and give up trying to reach the level of performance required to achieve a reward.

P4P is also said to crowd out individual initiative and discourage efforts to improve quality in situations where money is seen as substituting for professional and personal pride in performance. Sociologists report that once a level of extrinsic reward is reached intrinsic
rewards are highly motivating (Pfeffer, Cialdini, Hanna, & Knopoff, 1998). In some cases pushing monetary based P4P can be counterproductive. In fact, increased pay reduced physicians commitment to their hospital (Dukerich, Golden, & Shortell, 2002).

The health sector may attract individuals because of the intrinsic value associated with careers in this field. For example, many go into nursing not only because of the monetary compensation they will receive, but also because they wish to help people. This was observed in a study where a hospital lost doctors after P4P was implemented (Brian R. Golden & Sloan, 2008).

The question remains as to what will happen if P4P programs are deemed to have been so successful in changing provider behavior, that the P4P systems are not needed anymore. If one stops paying providers for their new behavior, will they revert to old ways? One classic sociological study tested this element in an industrial contest. The results showed that only when the experiment was actually ongoing did workers change their behavior. They reverted to old patterns once the study ended. There is some evidences that Pay for Performance does not work unless hospital or doctor behavior continues to be monitored. Sociologists know this as the Hawthorne Effect. Will it applies to P4P? More research is needed about post-P4P behavior.

A Distinctly Downside to Pay for Performance in the Health Sector

Experience with P4P has uncovered some specific negatives. Providers tend to treat a patient, even if the diagnosis is unclear or uncertain, if that treatment is rewarded under a P4P system (Jauhar, 2008). P4P increases disparities (Casalino et al., 2007). Many physicians do not like P4P, viewing it as intrusive, controlling, and punitive rather than as educational (Giraud-Roufast & Chabot, 2008).
The downside of P4P extends beyond clinical care considerations to management and structural considerations. It is expensive to set up, implement, and administer. Ongoing costs include monitoring and payment. It is thought to discourage innovation (Jauhar, 2008).

Applying Psychology and Sociology to Pay for Performance

While some aspects of P4P are structured and organized on the basis of knowledge from psychology and sociology, a good bit of it is purely intuitive. Not all relevant information from sociology and psychology has been integrated into P4P practice. For example, research suggests that the opportune moment to reward an individual for a specific behavior is immediately after that behavior is elicited. Rewards should be time-linked to the desired behavior with as short a gap as feasible between the two. Regarding whom to Pay for Performance, the literature suggests the individual whose behavior you wish to change should be the target. This might be a group or individual physician, hospital, patient. But P4P plans often target hospital and physician groups. Another consideration involves how to pay individuals and organization: with a carrot or stick? Should incentives be view as rewards or as punishment? More research is needed.

An additional point from sociology and psychology is that individuals do not all react to incentives in the same way. P4P should be aligned with an individuals’ personality if it is to achieve optimal results (Hamblin, 2008). Some personality types respond to one form of reward better than another. For example, in the health sector some physicians might respond to P4P that is linked to the ethic of providing healthcare to whole population rather than to personal financial incentives. This is almost never employed in P4P because of the complexity involved but it is well worth studying. Personal goals are similarly responsive to various P4P approaches. A provider that is self-interested might be open to different P4P formats than one that is more
altruistic. Research suggests that about 50% of the population is altruistic, 1/3rd self-interested, and the others mixed (Camerer & Fehr, 2006; Rosenau, 2006). This type of information might well be used to structure effective P4P programs. But applying different formats of P4P to different individuals in an organization has not been tried.

What Kind of Competition Does Pay for Performance Encourage?

P4P is based on assumptions about the nature of competition. How P4P programs are structured effects the type of competition that results when they are implemented. Much is known about the effects of various forms of competition on performance, but what is known has not been applied to the P4P programs in the health sector.

P4P often encourages the worst types of competition in terms of negative effects on performance – it is referred to in this literature as zero-sum competition (Rosenau, 2003). It amounts to taking away from those that do not meet P4P standards, and giving to those who do well on these criteria. It is budget predictable in the sense that no new funds are necessary. Medicare employs this form of P4P structure offering 10% bonuses to top performers and imposing a financial penalty of 10% reductions in pay to the bottom 10%. Across many economic and social sectors this form of competition has resulted in those in the bottom 10% becoming indifferent–many give up trying because they have no chance of “winning.” Those that do come out the best – the top 10% - in many cases are not doing anything that they were not already doing.

Another form of P4P competition sets a standard for excellence and offers a reward for all who attain this level of performance. This is sometimes called anti-anxiety, positive competition (Rosenau, 2003). It means that all can “win” if they meet the standard. None are punished for not meeting the standard. The downside of this structure for P4P is that it is
difficult to budget for anti-Anxiety, positive competition precisely because it is not possible to anticipate how many organizations or individuals while meet the standard and claim the reward. This approach was employed in Britain where a far greater number of doctors than anticipated meet the standard for excellence and earned a large bonus.

Anti-anxiety forms of positive competition make sense when the performance standard is within reach of all participants and when the desired performance criteria represent a cost-saving at the system level. Budget predictability does not matter if the goals are cost efficient and the goal is system-level improvement. In these cases the more you spend on P4P rewards, the more you save. In the end, the more “winners” that result from this type of P4P, the more savings accrue to the system. “The more you buy, the more you save” may not be possible in every situation but where it is, this form of P4P is advisable.

**Alternatives to Classic Pay for Performance**

P4P that employs rewards that are not financial can be very effective and do not have much of an impact on budgets. These P4P possibilities include public recognition, praise, attention to the fact that they (individuals or institutions) are “doing good.” The opportunity for more challenging work is another non-financial motivator (Dukerich et al., 2002; B. R. Golden, 2006, p. 2).

Finally, some say that P4P is directed at the wrong entity. They say that it makes no sense to reward doctors, if, for example, patients stop smoking. And another example, paying the doctor if the patient’s cholesterol is reduced misses the point and the same is true of weight loss and tobacco use. The reward must go to those who are really in control and this is the patient, not the provider (Volpp, Pauly, Loewenstein, & Bangsberg, 2009; Wharam & Sulmasy, 2009). The doctor’s ability to influence the patient’s life choices is never absolute. And it does make
sense to reward the patient in many situations. But research to test whether P4P for doctors or patients is the most effective when the performance metric concerns patient behavior.

**Conclusion**

P4P is a simple solution offered for very complex problems - improving the cost, quality and access to health care. Britain claims to have had excellent results with Pay-for-Performance for improving quality. But they are working on quality in ways that the US may not be able to imitate. About 30% of medical care in the US is for procedures and treatments that do not benefit the patient according to the recently released 2008 version of the Dartmouth Medical Atlas (Wennberg, Fisher, Skinner, & Bronner, 2007; Wennberg, O'Connor, Dale Collins, & Weinstein, 2007). If true, then P4P in the US may have to reward doctors for NOT performing medical procedures that don't need to be done, for NOT prescribing medications that are not appropriate for the patient, etc. But Pay for Performance alone is unlikely to correct this problem.

Critical to any conclusions about P4P in the health sector are the answers to the following question (B. R. Golden, 2006; Brian R. Golden & Sloan, 2008). Does P4P improve the health of patients? If so, how much? In the long term or the short term or both? Are studies valid across: all provider types, for all medical conditions, for all geographical areas or countries, for all cultures? In short, do research results generalize?

P4P is sometimes driven by preference, philosophy and ideology and it is often based on assumptions about human nature that are somewhat disturbing - i.e. that people, doctors, nurses, hospitals, educators, whatever, won't do their best because they believe in what they do and want to help others. In the end if we pay doctors and hospital to do something we should expect that
they will do their best and get the job done. Of course humankind is a mixed bag - some are good, some are not so good, but what does it do to our self-image if we start off with a negative assumption that all are not-so-good. Might targeting “outcomes” that are linked to values and indirectly influence behavior be more effective?

In US P4P cannot be recommended as “evidence based policy” at this point in time (Wodchis et al., 2007), because results are mixed and inconclusive. P4P is a blunt instrument, complicated by unexpected externalities. P4P has costs as well as benefits for the US and these have not yet been consciously and fairly assessed by policy makers, many of whom are predisposed to support these approaches in the absence of any evidence. If the goal is to improve quality, P4P may be of some help if it is appropriately structured. In terms of reducing costs – P4P may have some success but this is not a guaranteed. If structured inappropriately huge externalities can be expected. These are very difficult to factor into the equations that measure results.

Some lessons are clear from experience to date with P4P. Medicare learned and is now acting on the idea that medical errors should not be encouraged by reimbursement. This policy decision is, in some sense, an elementary form of P4P. It is important to view the medical care system as a whole, identify areas where change is needed, and pay for what needs to be improved. For example, in the US we need to pay doctors for agreeing to work weekends and after 5PM so the patients don't have to go to the emergency rooms to get routine treatment at these times. Dutch doctors are a good example for the US medical community in this respect. But it is very difficult to get US doctors and especially primary care physicians to do this after-hours and during the weekends. Finally, it is hard to avoid concluding that more research is needed because so many questions remain regarding P4P.
REFERENCES


