PIRU POLICY INNOVATION RESEARCH UNIT

Challenges of payment-for-performance in health care and other public services – design, implementation and evaluation

Mylene Lagarde, Michael Wright,
Julie Nossiter and Nicholas Mays
Policy Innovation Research Unit,
London School of Hygiene and Tropical Medicine

PIRU Publication
2013-4

For further details, please contact:

Mylene Lagarde
Policy Research Unit in Policy Innovation Research (PIRU)
Department of Health Services Research & Policy
London School of Hygiene and Tropical Medicine
15–17 Tavistock Place
London WC1H 9SH
Email: mylene.lagarde@lshtm.ac.uk
www.piru.ac.uk

PIRU POLICY INNOVATION RESEARCH UNIT

Challenges of payment-for-performance in health care and other public services – design, implementation and evaluation

Mylene Lagarde, Michael Wright, Julie Nossiter and Nicholas Mays
Policy Innovation Research Unit, London School of Hygiene and Tropical Medicine

August 2013

Funding
This work is funded by the Policy Research Programme of the Department of Health for England, via its core support for the Policy Innovation Research Unit. This is an independent report commissioned and funded by the Department of Health. The views expressed are not necessarily those of the Department.

Contents

Abstract 1
Introduction 2

Designing P4P schemes in public services 3
1. Who is paid? 4
2. What is paid for? 6
3. How is the payment structured? 9
4. Are unintended negative effects anticipated? 11
Conclusion 12

Implementing P4P schemes 12
What does it take to deploy and run a P4P scheme? 12
What can delay or hinder implementation of P4P? 15
What helps implementation of a pay for performance programme? 16
Re-visiting the original design – flexibility 18
Conclusion 18

Evaluating P4P schemes 19
What is the evidence on the impact of P4P in public services? 19
What are the technical challenges of deriving robust evidence on P4P? 20
What are the gaps in the evidence on P4P? 23

Conclusions 25
References 27

Abstract

Interest in mechanisms linking varying proportions of provider payments to achieving certain performance targets has recently grown in health care and other public services in an attempt to solve the cost-quality conundrum. Against the backdrop of this enthusiasm, it is important to review the challenges posed in the process of designing, implementing and evaluating of pay-for-performance (P4P) in health care and in other UK public services.

The design of P4P in the public sector is more difficult than in the private sector, principally because the production of public services usually involves the pursuit of several different objectives simultaneously that may be more or less easy to identify, measure and trade off. This tends to reduce the relevance and benefits of private sector financial incentives such as performance pay. Design is also complex because of the different dimensions along which any scheme can vary. The potential effects of a P4P scheme in the public sector depend on the interactions between: (1) who is rewarded: individuals, teams, or organisations; (2) what is rewarded and how is performance measured; and (3) how is the payment structured? There are several key questions in relation to the payment structure, such as how big is the conditional payment in comparison to unconditional payments? Are there rewards or penalties? If there are different targets, what weight is given to each?

Despite a relative lack of attention in the literature to issues of implementation, the way P4P programmes are implemented can influence the likelihood of success. At the outset, the context in which a P4P programme is being introduced can influence the success of implementation (e.g. whether providers are dependent on few or many funders). Effective communication of the rationale for the P4P programme is crucial to successful implementation. Appropriate monitoring systems are also key and the right balance needs to be found between providing accurate data that will avoid gaming and ensure that the programme is credible, and limiting additional administrative burden on providers. Lessons are starting to emerge on potential success factors, typically involving a long lead-in time to plan, test and reflect carefully on the different elements of the programme.

To date, the evaluation of P4P schemes in public services has produced limited and mixed evidence of their effects. In health care, many studies show either a positive, or a partially positive result, while others fail to demonstrate positive effects. There is very little evidence on the cost-effectiveness of schemes. There is also evidence of perverse effects of P4P. It is not possible to generalise about the effects associated with P4P since these are linked to the many ways in which the design of schemes can vary. The body of evidence is further limited by a frequent lack of political willingness and/or technical ability to introduce P4P on an experimental basis. As a result, control groups are often not available or highly imperfect. Other challenges to producing meaningful evidence of the effects of P4P include the limited time horizons of most evaluations, the complexity of the task of disentangling the effects of different components of programmes (which may also vary over time) and the need to look at non-incentivised as well as incentivised activities which can require more than routine data.

The jury is still out on whether the benefits of P4P schemes in public services exceed their costs. Indeed, it may not be wise to attempt to reach an overall verdict about P4P since the performance of each scheme is highly likely to be the product of the interaction between the detail of its design, the specifics of the setting, and the particulars of the provider and the service being delivered.

Introduction

To control rising health expenditure, use financial resources more efficiently and ensure the provision of adequate quantity and quality of services, public service payers have increasingly focused on altering payment mechanisms to create appropriate incentives for individuals, teams or organisations. Financial incentives can also be implemented for service users (e.g. patients) to encourage the adoption of desirable behaviours. However, this document focuses on incentives created for the supply (delivery) of public services. It discusses the design, implementation and evaluation of payment for performance (P4P) schemes, specifically to draw attention to the fact that all three activities are vital to the successful use of P4P.

There is a wide variety of provider payment mechanisms available to policy-makers and payers. Payment mechanisms can be linked to inputs, services or activities, population coverage or particular performance targets, and can differ in terms of how and when the payment is set and made. Each payment mechanism offers different types of incentives to providers, in theory, and generates both advantages and drawbacks in practice.

For example in primary health care, practices or individual practitioners can be paid by fee-for-service (FFS), capitation, salary, or a combination of these payment mechanisms. FFS reimburses health providers for a specific service or activity, which gives an incentive to increase services, even if unnecessary. Capitation provides a set payment for each person registered with a provider, offering incentives to contain costs, maintain patients' health or select out potentially costly patients thereby maximising surplus but risking under-service ('skimping') and/or low quality care. Although salary payments do not necessarily encourage over-servicing or patient selection, there is no incentive to increase effort or outputs. While some of these payment mechanisms may lead to efficient use of resources, none creates explicit incentives for high quality of care.

Most countries have moved away from using global, historic budgets to pay hospitals as this does not necessarily directly promote efficient use of financial resources (though a fixed budget should encourage some attempt to set spending priorities in a public system as demand generally outstrips supply), instead opting for prospective case-based payments linking the amount of money paid to patient diagnoses, characteristics and treatments. However, these payment mechanisms are not without their problems, as hospitals may actively seek patients with 'profitable' diagnoses to maximise income or reduce quality in order to make a surplus within the pre-set payment available per episode of care, assuming they know, or think they know, their cost structures. As patients with more severe conditions within each diagnostic category are likely to cost more to treat, they risk being avoided or under-served.

Interest in mixed payment mechanisms, in particular those in which part of the remuneration is dependent on achieving certain performance targets, has recently grown in health care and other public services in an attempt to solve the cost-quality conundrum. For instance, payment-for-performance (P4P), also termed 'Payment for results' or 'Payment by results'[1], is mentioned 15 times in the UK's 2011 Open Public Services White Paper. Various forms of P4P have been used across different sectors of policy in the UK in the last few years, most notably in welfare to work (the Work Programme), education (pre-school Sure Start centre pilots, teachers' pay) and health care (drug and alcohol recovery, smoking cessation and GP services).

Despite this enthusiasm, there is limited empirical work on P4P in public services. Schemes are typically justified on commonsense grounds (i.e. it must make sense to pay for the results, rather than the provision, of services). At a time when P4P schemes are being considered in a growing number of settings, including health care, it is timely to review the challenges linked to the design, implementation and evaluation of P4P, in health care and in other UK public services.

Designing P4P schemes in public services

Questions about the optimal design of payment schemes have their theoretical roots in the principal-agent model of behaviour [1]. A principal-agent relationship exists when a first party (the principal: typically the funding organisation) hires another party (the agent: a single worker, a team or an organisation such as a hospital) to complete a task (e.g. deliver services) on its behalf. As the effort undertaken by the agent is not necessarily observable by the principal (i.e. there is information asymmetry), the principal designs a contract linking rewards and effort (financial incentives) to ensure that the principal's objectives are followed by the agent.

The interest in private sector payment methods such as performance-related remuneration of individual staff has been growing in the public sector. This is despite the existence of several aspects in the production of public services that depart from those found in the private sector (which may reduce the relevance and benefits of private sector financial incentives).

First, the production of public services usually involves the pursuit of several simultaneous objectives that may be more or less easy to identify, measure and trade off. For example, hospitals should contribute to improving the long-term health of the population they serve, and they are also supposed to make sure the re-admission of patients due to complications following surgery is kept to a minimum. As some tasks are easier to measure than others, they might be the focus of financial incentives, at the expense of other less easily assessed activities.

Second, because the services are public in nature, several groups in society will be interested in their provision, cost and efficiency (e.g. governments, tax payers, citizens, direct beneficiaries, informal carers, relatives). Their objectives may directly conflict with the interests of others. For example, tax payers will generally want to keep down costs, but those who provide the service or directly benefit from it may

1 P4P is often called "payment by results" which can be confusing as Payment by Results is also the term used in the English NHS to refer to the prospective payment for hospital care based on activity, not on performance targets. P4P is also sometimes referred to as performance pay, performance-based funding or outcome-based contracting.

be more interested in high quality. Public funders will be typically interested in cost containment, equity of access and efficiency. The economics literature emphasises that in such situations where there are multiple principals and conflicting goals, financial incentives will be weaker than in the case of a single principal and a single goal (e.g. profit maximisation). Last, there is a broad literature supporting the idea that, given the nature of the "mission" or services provided by public sector organisations [2], public sector workers care about their job and how well they do it, and, even in the absence of explicit financial incentives, intrinsic motivation is likely to stimulate their efforts to carry out tasks that benefit others (students, patients, etc.). Indeed, in some circumstances, too much reliance on financial incentives could be de-motivating since it risks undermining providers' professional ethos and concern for maintaining a good reputation with their peers. Given this, creating effective incentives that will align these different objectives is likely to be more difficult in this type of setting than in the private sector (where extrinsic reward systems were first implemented and studied).

Keeping this in mind, to understand and anticipate the potential effects of a P4P scheme in the public sector, it is also important to identify the different dimensions along which any scheme can vary [3].

1. Who is paid?

Who should be rewarded for successful achievement of performance standards is a major consideration when designing a P4P scheme. To produce its most potent effect, the reward should be targeted where accountability for delivering improved performance lies. The standard model of agency assumes, in its simplest form, a contractual relationship between the principal and an agent who is uniquely and clearly accountable for the performance measured.

Paying individuals

If the desired outcome is the sole responsibility of an individual, then the P4P contract should be made with that individual. Drawing on the logic of performance pay in human resource management in the private sector [4, 5], this approach, which should be referred to as "performance pay", rewards individual health care providers for achieving pre-specified performance targets.

In health care, this might conceivably be the case in the context of single-handed practice primary care or office-based specialist care in some countries, if individual doctors are responsible for a given population, and their individual actions or performance can be accurately recorded and attributed to the outcomes achieved.

In England's NHS, the longstanding Clinical Excellence Award scheme is a type of performance-related pay. This scheme rewards NHS consultants who perform "over and above" what is expected of their role. Twelve levels of award exist with the lower-levels awarded locally, and the highest levels (bronze, silver, gold, and platinum) awarded nationally. These financial bonuses range approximately from

£3,000 to £76,000 for the highest "platinum" level[2]. Due to the opacity of the performance measure used and its weak link to actual performance (bonuses are awarded for five years and non-renewal of awards is rare), this scheme has been criticised repeatedly [6].

More generally, there have been two main criticisms addressed at individual performance pay schemes in public services. First, performance pay is likely to create high-powered incentives for individual workers, and consequently it may have a negative impact on workers' intrinsic motivation. The creation of explicit financial rewards through performance pay may signal a purely market relationship between the workers and the organisation, and thus dilute part of their original intrinsic motivation, thereby decreasing the effectiveness of the performance pay scheme [7]. Others argue that when the reward system is seen as controlling and limiting workers' autonomy, turning "knights" into "knaves" [8], intrinsic motivation is more likely to be undermined, with important implications for the degree of oversight and checking required to assure good performance [9]. However, to date there is limited evidence of a potential crowding out effect of intrinsic motivation.

A second criticism of individual level performance-related pay is that it fails to account for the fact that work in most organisations, particularly in the public services, relies on collaboration between workers to achieve the best outcomes. If this is the case, individual schemes are likely to undermine the combined efforts of individuals by limiting each member of staff to the pursuit of their own ends. Also, disentangling one person's contribution in a context where outcomes are met by team work is bound to be disputable and disputed. This difficulty with attribution is further complicated by the unpredictability of many health outcomes (discussed further below).

Paying teams

As mentioned above, very often in public services, the desired outcomes are not the sole responsibility of an individual, but rather the product of collaborative work in a team where all members work to provide services (either at different times, or by dividing complementary tasks between themselves). Sometimes, even if the steps taken to achieve the desired outcomes are taken by individuals, it is the case that performance data are only available at a more aggregate level than that at which production decisions take place. In both cases, the optimal solution to the agency problem is to reward the team rather than the individuals in the team.

Examples in health care include recent experiences in low- and middle-income countries where primary care facilities were allocated bonuses to be split up between the different staff members [10]. P4P schemes have been developed in these settings mostly to increase uptake of health care services, as well as to improve basic quality of services (through crude facility-level indicators such as availability of essential drugs). Schemes were designed at facility level mostly for monitoring reasons, as it would be infeasible to record individual performance.

2 http://bma.org.uk/practical-support-at-work/pay-fees-allowances/merit-awards-for-nhs-consultants/clinical-excellence-awards-consultant-scales

The main issue posed by team level incentives is that of "free-riding": if performance of individuals cannot be observed directly and all team members are rewarded (irrespective of their performance), some will shirk and rely on others to work towards the performance bonus. If teams are small and individuals see each others' work, peer pressure can counterbalance the free-riding issue somewhat. While team incentive schemes seem to overcome the problem raised by lack of performance monitoring at the individual level, this issue may re-enter through the back door when it comes to knowing how team incentives are distributed to different individuals. For example, perceived equitable distribution of performance-related income to individual team members is likely to encourage team cohesion and cooperation [3]. On the other hand, if the rewards in a team are not distributed relative to the perceived effort of individual staff, higher performing individuals could become de-motivated over time as they see colleagues 'free-riding' on their efforts.

Paying organisations

Finally, central or local public authorities can decide to commission the delivery of public services from an organisation (e.g. hospital, general practice, school, employment service provider).

The assumption is that since the terms of the payment are partly or completely conditional on achieving performance targets, the contract provides incentives to the organisation and its management to innovate and make the necessary adjustments to its production processes to achieve a particular set of results. In health care, contracts are frequently signed with a medical group or hospital. When these contracts include a P4P clause, these organisations, in turn, have to decide whether or not to create explicit financial incentives for the individuals or teams within the organisation accountable for the performance targets [11]. For example, in the recent P4P pilot for NHS hospitals in North West England [12], the performance payments to the organisations did not translate into bonuses for individuals, but instead any additional income went to clinical teams to be spent on improving clinical care in their service area. In cases where P4P contracts with organisations do not translate into individual or team incentives, the crowding-out effects on intrinsic motivation discussed above may not necessarily apply.

2. What is paid for?

In the simplest principal-agent framework, incentives are created by linking (all or part of) the agent's compensation to the desirable performance measure. When performance is perfectly observable, as in the case of a salesman's output, the value of that output provides a perfect indicator of the agent's effort, and paying an individual the full value of her/his output induces an optimal level of effort. However, the most easily measured performance indicators might not necessarily be the most important from a policy viewpoint. What can be measured thus risks taking priority over what is of most importance during the design of a functioning P4P scheme.

Defining the objective of the scheme

As noted in the introduction, one of the complexities of paying for public services lies in the co-existence of different principals, who might have diverging interests. The financial incentives created by a P4P scheme are a way to align the behaviours of the providers with one objective that a principal wishes to see prioritized. While they can be geared to focus providers' attention on improving efficiency of resource use, in health care, P4P schemes have been mostly used, so far, to incentivise providers to improve quality of service, with the ultimate objective of achieving better health outcomes, but there is no reason, in principle, why other objectives such as responsiveness or efficiency could not be the focus of schemes [3].

Choosing the performance measure(s)

Irrespective of the objective of the P4P, the principle is the same in each case – payment to the health care provider is made conditional on meeting a measurable result that will allow the payer to achieve the pre-specified objective (improving quality, efficiency, health, etc.).

Obviously, performance measures chosen should be associated more or less directly with the desired objective. But as noticed in the introduction, some objectives in public services can be difficult to define and therefore difficult to measure. Providing a 'good education' (for a school) is harder to define than, say, the adequate collection of garbage. In addition, it takes much longer to determine whether a 'good education' has been delivered than whether the garbage has been adequately collected. This is why very often performance measures are identified that are associated with the final desired objective, but do not necessarily capture that objective directly. Three types of measures can be identified that assess different elements of the production process:

- Input measures that describe the appropriate environment and resources used in the delivery of a service;
- Process measures that consider the steps taken that are likely to achieve a particular result or specific activities using inputs;
- Output measures that look at the results of the delivery system. These can be 'final' results (e.g. health improvements) or intermediate results (e.g. better glycaemic control in diabetes).

In health care, P4P is often linked to measurable process indicators, in particular, the completion of activities that are widely regarded as highly likely to contribute to good quality of care and/or improved health outcomes (see table 1). For example, following validated disease management protocols is generally associated with better health outcomes than not. Process measures have the advantages of being relatively straightforward to monitor and of linking relatively directly to providers' efforts and the related costs of delivering the desired outcomes.

This classification can be used to think about the types of measures that can be chosen in a P4P in health care.

Table 1 Examples of performance measures used in P4P schemes in health care

	Measure		
Objective	Input measures	Process measures	Outcome measures
Improvement in quality of care	IT system in place, opening hours widened	Proportion of patients treated according to a specific clinical pathway	Patient-reported outcome measures, Health-related quality of life
Improvement in health	Uptake of preventive health services (e.g. immunisations, deliveries with skilled birth attendants)	Proportion of target population provided with particular preventive information (e.g. smoking cessation options available)	Mortality, hospital readmission rate

To maximise their chances of generating powerful incentives, the human resources literature recommends that targets be SMART: specific, measurable, achievable, relevant and time-limited. In public services, such targets may be hard to identify and, as a result, the measures adopted may be very imprecise. There is often some concern that the chosen performance measures are poor proxies for the ultimately desirable, but hard-to-measure objective (e.g. quality of care) and may simply increase activity related to straightforward measurable tasks over more valuable but harder to measure tasks [13]. If performance measures are too imprecise, they will not convey relevant signals to providers about the ultimate objective of the scheme and it is likely that these objectives will only partly be met.

Another problem in the production of public services is that the desired objective (and even performance measures) may also be dependent on a (random) component beyond the agent's control. For example, results in exams will depend on teachers' performance as well as students' efforts and motivation; use of health services will depend on providers' efforts, but also on patients' health seeking decisions and underlying health status. In these situations, it can become impossible for the principal to identify a performance measure that rightly reflects the effort of an employee or organisation. As a consequence, linking rewards to the meeting of performance targets in this situation does not provide effective incentives, may lead to 'gaming' and imposes unnecessary risk on providers. This is why performance measures generally need to be risk-adjusted in order to control for such variations in the populations served by different providers. Although risk adjustment may lead to fairer and more accurate performance measures, the methods used can also be controversial in themselves, and they add a further level of complexity to the signals sent by payment for performance schemes to providers [13, 14].

3. How is the payment structured?

How much?

The size of the financial reward to be received upon achievement of a desired target in a P4P scheme must be determined. Economic theory suggests that the size of the reward should be proportional to the effort required to obtain the performance improvement, but in reality this relationship is often unclear and the optimal size of the reward remains uncertain [15]. In theory, larger incentives are associated with greater improvements in performance compared with smaller incentives [16], with the latter risking being insufficient to incentivise change. However, as the size of the reward increases, so too does the risk of adverse consequences (see below).

In the context of health care, some have challenged the theoretical framework above, and have argued that the size of the reward should be proportionate to the likely health gain and thence the value associated with achieving each performance target, rather than the amount of effort required by the provider to reach that target [17]. While this is attractive in theory, in most cases P4P schemes have to be negotiated with providers (e.g. as part of the industrial relations bargaining between professional groups and public funders) who are far more likely to prefer effort-related performance payments than value-related performance payments since the likely relative costs of achieving the necessary target are built into the former but not the latter.

In most public sector organisations, P4P schemes provide financial incentives that complement other remuneration mechanisms, such as salaries or budgets. Not only should the absolute size of the P4P reward be considered, its relative importance in relation to other sources of payment will also determine whether it is a key incentive for motivating agents to undertake the necessary effort or not [14]. For instance, in the recent 'Advancing Quality' P4P pilot for NHS hospitals in the North-West of England, [12] hospitals could earn a bonus equal to 2% to 4% of the revenue they received under the national activity-based tariff. By contrast, general medical practices under the NHS general practice quality and outcomes framework (QoF) can earn a bonus representing an additional 25% of their income on average [18].

There is a growing interest in performance contracts with private (not-for-profit) providers that put a lot of their remuneration at risk, sometimes up to 100% [19, 20]. While proponents argue that such schemes will promote innovation in service delivery, such schemes also raise questions about the ultimate responsibility of governments if privately contracted providers fail to meet their targets and/or become insolvent attempting to meet demanding targets.

Absolute or relative targets?

There are multiple ways in which the P4P contract can define the payment conditionality. Most P4P schemes condition payments on achieving targets for key performance measures. The advantage of using absolute targets is that there is no uncertainty as to whether or not a particular standard has been met.

However, the suitability of absolute standards to incentivise both exceptional and poor performers has been questioned [21]. Those already meeting the threshold or performing above it will be rewarded, with little incentive for further improvement, whilst those far below will require a significant incentive to encourage the efforts required to reach the target and may give up entirely [11].

When a principal contracts several agents to provide services on his/her behalf, P4P contracts can be designed to reward only the top performers (e.g. the best 10%). This system, called a 'tournament' in the economics literature, aims to promote competition and encourage all agents to increase their efforts, while the principal only has to reward some of the participants. This design was used in the 'Advancing Quality' P4P scheme implemented in the North-West of England [12]. However, only rewarding relatively high performers presents several potential drawbacks. Such systems may create uncertainty for providers and payers as payment depends on the performance of all providers rather than each provider's own performance [22]. As a result, some might be discouraged and not even try to make the necessary efforts. Furthermore, rewarding relative performance may further exacerbate the performance gap between high and low performers [22]. This might be particularly undesirable in the provision of public services, where concerns for equal standards of service quality are prominent. Finally, competition may erode more collegial behaviours, thereby limiting collaboration and sharing of best practice which could adversely affect overall performance [22]. However, contrary to these predictions based on economic theory, in the pilot in the North-West of England, staff from all participating hospitals adopted a collaborative approach and met regularly to discuss similar problems and share lessons about solutions implemented to tackle them. A historic collaborative culture in the NHS appeared to trump more recent performance incentives.

How many targets?

If the target is set too low, high performers will meet the target easily and so funds will be wasted on overcompensation. Conversely, if the target is too high, poor performers will be discouraged [15]. To address some of these issues, it is possible to set a number of thresholds, and/or a sliding scale as in the Quality and Outcome Framework (QoF). This system pays increasing amounts of money with higher rates of achievement over a minimum level. For example, in the QoF, payments are typically triggered on reaching a lower threshold (e.g. vaccination rate), with increasing payments up a sliding scale until an upper threshold is reached.

Rewards or penalties?

Instead of rewarding a desired output, P4P schemes can penalise poor performance either by fining low performers or by withholding a proportion of funds until they achieve a predefined level of performance. While, in theory, withholding payments for under-performance may induce a greater behavioural response than the promise of a financial reward, due to loss aversion [23], there is limited evidence on the relative effectiveness of penalties versus bonuses in P4P schemes [24].

4. Are unintended negative effects anticipated?

As they sometimes provide sharp financial incentives (often deliberately), even when they are well designed, P4P schemes may create some unwanted consequences.

Such P4P schemes are likely to encourage providers to try and 'game' the system to gain the financial rewards without the effort. This is particularly an issue when monitoring individuals' performance is not straightforward and relies entirely on a reporting system that can be manipulated by the very individuals who benefit from the rewards. Analysis of a cheating system developed by teachers and principals in Atlanta public schools demonstrated that the observed frequency of cheating strongly correlated with minor changes in incentives [25]. The authors concluded that *"high-powered incentive systems, especially those with bright line rules, may induce unexpected behavioural distortions such as cheating"*. This problem will be particularly acute when performance measures are imprecise and do not rely entirely on the agent's efforts to be achieved.

Another key insight from the economic agency model [26] is that workers will tend to focus their efforts on incentivised tasks ("tunnel vision"), rather than on tasks that are not so obviously rewarded (unless countervailing management efforts are put in place). Therefore, P4P schemes can be expected to have a detrimental impact on the achievement of unrewarded outcomes or efforts. To avoid such problems, it is often recommended that P4P schemes include safe-guarding conditions. For example, the payments might only be triggered if a certain level of performance is also achieved for important (non-incentivised) activities, or a certain percentage of remuneration could be withheld to be allocated ex post to performance on a previously un-incentivized area (i.e. building a degree of uncertainty into the process to keep providers focusing on a reasonably wide range of performance goals).

Another possible detrimental aspect of P4P schemes introduced amongst a group of providers is that they might increase existing inequalities in outcomes. Providers serving more affluent populations might achieve the (highest) P4P targets more easily and quickly than providers serving poorer and/or hard-to-reach populations, thereby further widening gaps in resources (as well as performance). This problem will be partly avoided if the schemes avoid rewarding only the best performers (as in a tournament style scheme), but instead offer gradually rising rewards to providers, starting at a relatively low level to encourage poor performers to take part. Another possibility is to take account of starting points and reward each provider on its relative performance improvement.

Finally, if providers are paid the same amount of money to achieve a level of performance that requires different levels of effort, they will tend to focus on the easier targets unless the performance payments are weighted for the expected effort required. In the context of health care, it might mean that providers will 'cherry-pick' patients and focus their efforts on those patients whose behaviours they are more likely to change. For example, increasing the uptake of immunisation may be much harder in some population groups, and might require more effort

from providers. To avoid such problems (also termed 'cream-skimming') , it is important to adapt rewards to reflect the efforts needed to attain particular objectives. This is when risk-adjusted or effort-adjusted targets are required.

Conclusion

This brief summary of P4P design issues shows the wide range of parameters on which schemes can vary and establishes a priori the trade-offs involved in developing any scheme. All will generate some downsides. In a systematic review of the evidence concerning P4P programmes in health care outside the United States, Eijkenaar [27] noted that the heterogeneity in design characteristics seemed to be partly a consequence of contextual differences, but also that design choices seemed to have been made arbitrarily. The author questioned the extent to which those who had designed schemes had had sufficient knowledge of the range of specific design choices, their potential influence on results and effective strategies to mitigate undesired provider behaviour. Of course, it is often the case that as the design (and implementation) of P4P schemes are shaped by industrial relations negotiations (e.g. on teachers' pay or general practice remuneration in the UK), they will involve compromises, not all of which will be desirable from the payer's or users' perspective. This means that there are likely to be compromises between different design principles (e.g. providers may take on less financial risk than funders would prefer, but this might reduce their risk of failure).

Implementing P4P schemes

In this section, we briefly cover some of the issues that may arise when implementing P4P programmes. We try to reflect some of the messages from the research literature. However, there is a relative lack of attention in the literature to issues of implementation as against design and subsequent evaluation. Often, implementation and design issues are inadequately differentiated. We have tried to reflect specifically on the problems that policy-makers should bear in mind when preparing the deployment plan of a P4P scheme, and what challenges to its implementation they could expect to encounter. In doing so, we have drawn on a workshop on P4P organised by PIRU in February 2013 (www.piru.ac.uk/events/payment-for-performance.html).

This section should provide some useful guidelines to implementers, highlighting the typical obstacles to implementation, and suggesting some ways to overcome them.

What does it take to deploy and run a P4P scheme?

There is a surprising lack of studies or even commentary analysing the work involved, and challenges of, implementing and running P4P schemes. Yet, the administrative challenges associated with deploying, managing and refining P4P schemes are bound to be significant, in particular, in terms of communicating the scheme to potential participants and setting up information systems to track performance validly.

A number of the issues for implementation will arise from the choices made in relation to the four key design characteristics (choice of performance measure, payment conditionality, choice of beneficiary and payment characteristic) discussed in the previous section. These choices may alter the acceptability of the scheme to different stakeholders and may also influence the focus of an implementation strategy.

Context

At the outset, it is important to recognise the context in which a P4P programme is being introduced. There may be other factors that influence the success of implementation (such as political, economic, or structural factors). These factors may be not be directly under the control of the implementer, but need to be considered in the design and implementation, such as the importance of establishing the level of support among providers before initiation. For instance, a single payer system (such as the UK's NHS) may be more successful in implementing a P4P programme than a fragmented system with multiple providers and incentives (such as the multiple insurer and provider market in the US). According to economic theory, the clearer the signal created by an incentive, the more individuals or institutions are likely to respond to it. Therefore, in a market where providers are paid by multiple insurers (each one designing their own reimbursement mechanism), the signals created by P4P schemes are more likely to be lost in a multitude of other signals than in a more unified single payer system.

Communication

Effective communication of the rationale for the P4P programme is crucial to implementation. A communication plan is necessary to ensure providers' awareness and understanding of the scheme, and ultimately to gain their acceptance and participation. This support is most effectively obtained during the design phase, as it allows providers to 'own' some of the design choices discussed earlier. Those engaged in the design stage, will have greater understanding of the purpose of the scheme and may be able to be co-opted as local advocates for the programme during implementation.

Because P4P touches upon sensitive issues (e.g. professional discretion and autonomy, remuneration and external assessment of the quality of service provided), its successful implementation is difficult without obtaining the initial, in principle, support of service providers. Since P4P systems are sometimes relatively complex and mostly fairly new, transparent policies and procedures must be defined. Efforts need to be taken to ensure that providers are aware of the nature and extent of the links between their actions and the incentives. Lack of transparency in the P4P system or lack of understanding of its functioning can lead to provider frustration and eventually to loss of trust in the payer [28], thereby contributing to the failure of the scheme entirely [29]. To ensure the success of any scheme, providers should be thoroughly briefed so that they understand the rationale for the performance being rewarded, the specifics of the scheme, and what actions are expected of them to attain the potential

rewards. Providers are more likely to engage with the P4P programme if the programme's incentives are aligned with their existing professional motivations.

Monitoring and auditing

Data gathering and reporting is central to pay for performance systems. There is often a trade-off between using available data (cheaper and more recognisable to providers, but which may have insufficient depth) and gathering data specific to the P4P programme (and the additional costs associated with this). Where feasible, existing data sources and processes of data collection should be used. However, in practice, the introduction of P4P schemes in health care is likely to require the development and introduction of specific, sometimes quite complex, support systems, such as new software packages. In turn, this triggers the need to train and support providers and staff in funding and providing organisations in the use of these new tools.

Monitoring performance requires the collection and analysis of suitably accurate performance data. The data necessary to set up and evaluate P4P schemes are usually significant. Accurate and suitably risk-adjusted data are necessary to avoid gaming (the manipulation of results to maximise performance) and the loss of credibility of a scheme. Existing information systems may not provide sufficiently accurate data. For example, P4P in health care may require additional clinical diagnostic and outcome information beyond that held by a particular provider. Raw data often require risk adjustment for confounding effects of external factors that may account for differences in performance (such as differences in case-mix). Support systems are also needed to provide feedback, so that decision-makers and provider staff know, on a relatively regular basis, how they are performing. Yet, for a P4P programme to be successful, it must not be overly complicated or expensive to administer otherwise it is unlikely to be cost-effective to introduce. Processes must be well structured and as simplified as possible, so that benefits exceed costs of implementation.

Agency theory suggests that independent auditing of performance is required under any payment arrangement that seeks to reward specific behaviours. Information collection, an essential component of P4P schemes, often requires costly up-front investment in IT systems, training and software. A key question is who will bear the costs of this monitoring, and to what extent the payer will seek to shift some of the costs onto the providers as part of the P4P contract. The fixed and/or running costs of setting up new information systems can be substantial for providers, in particular, small ones. In any case, as obtaining adequate performance measure is key to the successful functioning of any P4P scheme, it is in the interest of payers to make sure that adequate resources are dedicated to P4P administration to avoid measurement error, erroneous algorithm calculations, payment delays or inaccurate payments, all of which could harm the credibility of the programme.

What can delay or hinder implementation of P4P?

The literature on implementation of innovative medical treatment identifies a large number of barriers to implementation that prevent or slow down implementation of new treatments [30]. There is no similar research to draw upon for P4P schemes, but there are similarities between the two processes of change as a result of which it is possible to list a number of issues that may trigger negative reactions or opposition that can delay or inhibit P4P implementation:

- Lack of understanding of the objectives or functioning of the scheme, which may undermine the original goals or design features of the scheme. An effective communication strategy and adequate time (and money) are important to avoid this problem.

- Disagreement with the principles of P4P: concerns may exist over the choice of rewarded measures; the ability of the scheme to achieve improvement in quality of care, in particular, and health outcomes, in general; or its over-emphasis on short-term "box-checking" actions that are unrelated to good clinical care.

- Resistance to a scheme that requires additional reporting processes which can be judged to be cumbersome, time-consuming, or difficult to follow. Having too many measures can also hinder implementation, and having measures which are not recognised as being within the providers' control can also induce resistance.

- Inability of the P4P scheme to account for some specific constraints, such as geographic, linguistic or financial barriers faced by patients, which limit providers' capacity for quality improvement as measured by the P4P scheme. As a result, tensions may arise from local stake-holders who may reject the legitimacy of a scheme that does not account well enough for local particularities (such as failing to recognise, either through the reward or the performance measures, that the same achievements might require differential effort in different contexts). There is a tension between localism (and having local measures which may be seen as more relevant) and consistency (simpler to implement but seen as less relevant and thus less likely to succeed). Local measures may be useful to gain initial acceptance, but may be less useful to the monitoring of a P4P programme, its transferability or its cost. Overall, the best option might be to try to adapt a scheme developed and tested elsewhere to local priorities and context.

- Specific financial or administrative challenges might emerge during implementation that had not been anticipated by designers. For example, when P4P schemes put a non-negligible fraction of the provider's budget at risk, poor performance might jeopardise the future financial sustainability of providers, and thereby the capacity of providers to deliver services. Conversely, providers, particularly early volunteers, might be over-optimistic about their performance when they commit to taking part in a programme, which could lead them into financial difficulties later on.

- When P4P targets organisations rather than individuals, complex broader factors such as organisational climate[3] are likely to influence individuals' behaviour. In particular, if entrenched processes need to be altered to meet P4P targets, financial incentives may be regarded by some targeted users as a poor "values fit" [31] which might hinder their effectiveness. For example, if, as a result of the introduction of a P4P scheme, an organisation changed the management of patients suffering from chronic conditions to a more team-based approach, the change could be perceived as a "bad fit" by those physicians who hold their autonomy as a preeminent organizational value, but may be perceived as a "good fit" by nurses who believe collaboration in the treatment process is a preeminent value [31].

- Finally, practitioners might simply disagree with a remuneration system that assumes that outcomes should improve if they take specific actions, but ignores the role of chance in delivering good (or bad) outcomes and underestimates the difficulty of measuring the impact of their efforts. The assessment of 'results' needs, as far as possible, to try to take into account random variation (e.g. from period to period) just as in a research study.

More evidence is needed on the relative importance and frequency of these issues, and whether they are more or less likely to be encountered in different public services and particular types of schemes. One might also hypothesise that high status health care professionals are more likely to succeed in resisting and altering P4P schemes in their favour than, for example, classroom assistants in schools.

What helps implementation of a pay for performance programme?

There is increasing experience with these programmes and a number of potential success factors are beginning to be identified:

- Having a long lead-in time is likely to improve success [32]. This allows more time for participants to become familiar with the programme, their role, and its associated rules and rewards. This may represent a period of time after initial design, but before initial implementation of the programme.

- Testing the feasibility of a P4P programme, its acceptance and its potential adverse effects is also highly likely to be important before full implementation. Piloting of this type should permit some flexible redesign to the programme before full implementation, when some effects of different contexts are more likely to be known (although some flexibility will still need to be retained even after full implementation).

- Along with piloting, implementation may be easier if using a programme which has been successfully implemented elsewhere. A previously used ('tried and tested') programme may have resolved many of the design and implementation issues, but consideration must still be given to adapting it to the new context

3 In an organisation, this is closely linked to the work culture, or properties of the work environment directly or indirectly perceived by employees.

so that an imported programme is seen to have local relevance. There is a careful balance to be struck between wholesale adoption and appropriate modification to existing programmes that have been tried elsewhere.

- Using data collected to improve performance (not just to allocate payment). Information systems collecting routine data for payment can also be used to highlight variations in performance over time and make comparisons with similar providers. Providing feedback to participants to inform their learning stresses the quality improvement aspect of the programme rather than focusing on the payment function or the probity of the scheme.

- Having a dedicated support system and infrastructure in order to rapidly provide assistance and identify any issues at an early stage. Dealing with problems early and effectively should help reduce the risk of early disenchantment.

- Learning from mistakes and remaining flexible about programme design to adapt it in response to changing circumstances.

- While competition between providers can be used as a driver up to a point, it is important also to encourage collaboration between providers to drive improvement. By sharing understanding of success factors, all providers may be more motivated to improve their performance. In the AQ pilot scheme in the North-West of England, best performers successfully shared improvement tips with other participating hospitals, despite the competitive nature of the scheme.

- Using positive language appears helpful in promoting motivation (for instance, presenting early challenges as an 'opportunity for improvement' rather than signs of incipient 'failure' may decrease the potential for early de-motivation and disengagement on the part of some participants).

- Encouraging the use of locally developed performance measures as a way of increasing engagement with providers. Using locally developed measures may improve initial acceptability to providers, but adds ongoing cost and complexity, and may reduce transferability of results. It is important to distinguish between the potential need to tailor P4P schemes to local priorities and allowing the measures to be designed locally without expert input.

- Adapting programmes in response to improvements or problems. The UK's QOF program reviews measures annually and allows retirement of measures felt no longer to be useful. This allows the introduction of new measures which may have increased importance.

- Recognising that behaviour change is very difficult to institutionalise. Simple 'pump priming' of the system is unlikely to be successful. An ongoing commitment in money and time is needed along with high level support from management.

- Other regulatory measures may be necessary to encourage involvement. Linking participation in P4P programmes to provider registration or practice

accreditation may encourage uptake (but may reduce acceptance of a programme by removing voluntarism which has been seen as important by some participants).

Re-visiting the original design – flexibility

As a result of the potential problems in implementation outlined above, the design of P4P schemes may need to be re-visited during implementation. (This is particularly likely to arise if individuals involved in the implementation of the scheme were not involved in its design, and might only realise what the incentives entail for them once the scheme is in operation). For example, concerns might arise about P4P programmes incentivising providers if schemes do not adequately reflect differences in patient mix. As discussed earlier, providers treating patients at higher risk or with more complex pathologies may feel that it is unfair to be judged only on process or outcome measures on which they will systematically score worse than others despite making equal efforts to achieve high quality care. Although P4P schemes might be designed to account for patient differences, these adjustments may still be controversial or unsatisfactory.

Refusal to address providers' concerns may threaten the legitimacy and sustainability of any incentive programme. Failure to account for providers' concerns may result in loss of faith or trust in the payers [33]. This breakdown in the relationship might cause damage which might not be repairable simply by a change in the design of the incentive structure. Refusal to take into account providers' concerns may also prompt gaming behaviours [34], or may prompt providers to leave the scheme (when participation is voluntary), which could have severe implications for some service users.

Conclusion

If P4P schemes are designed and implemented by policy makers or managers with little consultation with practitioners and service providers, they are likely to run into implementation difficulties. The issues highlighted above call for a continuous and flexible monitoring, revision, and improvement of P4P programmes after initial implementation, which should increase motivation and commitment of implementers and target users [35].

Evaluating P4P schemes

In this section, we highlight the main findings and gaps in the evidence on the costs and impacts of P4P schemes, and reflect on the need for, and challenges involved in, carrying out evaluations of P4P.

What is the evidence on the impact of P4P in public services?

Reviews evaluating the impact of different forms of performance payment in public services[4] in the UK and other countries highlight the lack of robust, convincing evidence as a consequence of *"the inability or unwillingness of governments to carry out experiments within the public sector [where] changes are either introduced for all service deliverers or all eligible recipients, or they are not introduced at all"* [36].

In education, evidence for the incentivising effect of P4P schemes on teachers is mixed. Robust experimental evidence supports the effectiveness of P4P targeting teachers to improve students' outcomes in India [37], Kenya [38] and Israel [39, 40]. However, recent experimental evidence from two P4P schemes in the US (one rewarding teachers individually [41], the other one rewarding schools [42]) shows that P4P programmes had no or little impact on pupils' performance, at least in the short-term. Yet, non-experimental studies from the US and the UK (relying on longitudinal routine data) report positive effects on student outcomes [43-46]. It is highly likely that a complex range of contextual and population features shape these different results.

Evidence from other public services forms an even less coherent body of work due to even larger differences in schemes, populations and settings. A team-based[5] incentive programme in the UK rewarding Jobcentre Plus employees for the number of job placements and the quality of their services had no impact on average [47], but there was significant heterogeneity in the performance of offices in achieving job placements (with smaller offices performing better than larger ones). In a randomised trial of team-level P4P schemes at HM Customs and Excise, there was evidence that incentives increased team performance, partly resulting from a management decision to reallocate efficient workers to the incentivised tasks [48].

In health care, several systematic reviews have been carried out on the effects of P4P, a number of which assessed the strength of the available evidence [3, 49-56]. There is some consensus emerging on the following points:

- *The evidence on P4P in health care is still mixed*, with many studies showing either a positive, or a partially positive result, with others failing to demonstrate positive effects on clinical performance. As noted in a recent Cochrane review [54], most of the studies in this field do not employ experimental designs[6]. As a result, in most cases, it is not clear whether the effects observed are a direct consequence of the introduction of P4P, or confounded by other factors (e.g. self-selection of providers or patients, other parallel interventions, etc.).

4 Usually defined publicly financed services, provided by publicly, privately or Third Sector owned providers.

5 The performance was assessed at the district level, each district being composed of a varying number of offices, which are the operational level where teams of workers interact with clients.

6 And those that do may be highly atypical.

- *There is some evidence of perverse effects of P4P programmes.* For example, non-incentivised activities may be affected negatively by P4P [57], and some providers may try to 'game' the system [58]. Yet such evidence remains limited, and it is unclear whether this behaviour stems from more than remediable loopholes in the design of schemes.

- As expected, *the effects associated with P4P programmes tend to be directly linked to the design characteristics of schemes*. For example, improvements in the quality of care slowed once maximum threshold targets had been reached (above which no bonus payments were made) in the UK NHS general practice Quality and Outcomes Framework (QoF), suggesting a ceiling effect [18]. In a study testing the impact of rewarding clinics with financial incentives for reaching a smoking cessation target, researchers suggested that the small size of the performance payment resulted in limited or no effect [59]. Overall, the very large number of different dimensions on which schemes can vary in their design is likely to explain the number of ways P4P can vary in their effects, even before taking population and contextual factors into account.

What are the technical challenges of deriving robust evidence on P4P?

The lack of robust evidence on the causal link between P4P and various outcomes partly derives from an inability or unwillingness of policy makers and implementers to introduce P4P schemes on an experimental basis, which would allow for the identification of a control group against which to compare the effects of introducing the intervention. In 2006, a review of P4P studies in health care demonstrated that nearly half of the eligible studies did not include a control group or compare quality indicators at baseline [50].

In fact, P4P schemes are often either introduced among those who volunteer to participate in pilot programmes, or at full scale across the board. Both cases present threats to the internal validity of evaluation studies. One response is to encourage scheme designers to work closely with evaluators from the earliest stages so as to build considerations of evaluation into not only the implementation process for schemes, but also into the designs of schemes themselves so as to increase their ability to produce meaningful findings of wide relevance. However, for this to be successful, the evaluators must be given the same status as policy officials, service commissioners, managers, professionals, users and their representatives in any process of 'co-design'. This is almost never the case.

When reforms are implemented nationwide, there is no obvious comparison group to compare the changes in outcomes observed before and after the introduction of the P4P scheme. Without a counterfactual, it is difficult to ascertain whether the changes observed are the result of the introduction of the new scheme, or would have happened anyway. Many changes over time can affect the settings where P4P is introduced and confound the estimated impact.

P4P in health care is particularly vulnerable to such problems, as unrelated changes in medical technology, clinical practice or payment methods occur continuously and are likely to affect outcomes of interest positively or negatively. Even with longitudinal data, where changes over time might be taken into account, it can be difficult to know whether the estimated impact is the true effect, or whether it is under- or over-estimated due to other concurrent factors, whose effects on the outcome of interest (in particular, health outcomes) might be dynamic or lagged.

To address these issues, researchers often seek to construct counterfactuals even in the absence of obvious ones. In some cases, it is possible to use the fact that some conditions are incentivised and other similar conditions are not. Comparing the evolution of outcomes in both groups, although imperfect, provides some confidence that the changes observed in the incentivised conditions can be attributed to the intervention. This was the approach in a recent evaluation of the Best Practice Tariffs in the English NHS[7] where changes over time in outcomes relating to an incentivised treatment were compared to the same indicators for other similar but non-incentivised treatments [60].

When P4P schemes are implemented at the national level, but where the choice of incentivised activities varies across areas, evaluations can similarly use the providers who have not chosen to have a given activity incentivised as comparators. However such comparisons between places that have or have not focused on a particular outcome are subject to high risk of selection bias even when attempts are made to match sites on the basis of seemingly relevant characteristics.

Selection bias is particularly problematic for pilot programmes with volunteer providers, where those who agree to participate might not be representative of their sector (for example, better performing or more motivated providers might be more willing to take part in a novel or more demanding scheme). To address some of these problems, evaluators typically attempt to match the comparison group as closely as possible with intervention beneficiaries [61]. However, even when matching on observable characteristics, differences in unobservable characteristics may remain between intervention and control groups, and lessons learnt from a small group of self-selected volunteers will never be completely applicable to all the other providers in a given sector or country. In the context of health care markets, provider competition can generate spill-over effects in control groups. For example, if one hospital demonstrates improvement, market forces are likely to drive neighbouring hospitals to improve as well. Although this is potentially desirable, in the context of an evaluation, it can lead to an underestimation of the true effect of P4P in the intervention group.

Even if the thorny problem of identifying a counterfactual, or relevant comparison, is (partly) solved, other shortcomings and challenges may arise during the evaluation of P4P programmes:

- *Establishing and maintaining agreement as to the goals of P4P schemes.* Schemes have been advocated to reduce costs, generate 'cashable savings', improve efficiency, improve quality and outcomes, transfer financial and

7 A P4P scheme reimbursing hospitals at a higher rate if specific protocols are followed: incentivising day case treatment for cholecystectomy, paying for best practice for stroke and hip fracture, and a streamlined care pathway for cataracts.

service risk to the provider, particularly the private sector and encourage innovative approaches to service delivery. Evaluating consequences under each of these headings is demanding in itself, never mind trying to look at each simultaneously, especially if their relative salience changes over time.

- *The time period over which the effects of the scheme are evaluated might be too limited to produce reliable results.* This can be a particular risk in the case of a pilot scheme. For practical and/or political reasons evaluators frequently have to complete their research in a short timeframe. Short follow-up periods also tend to constrain the scope of outcomes to be evaluated (e.g. some health outcomes may only become apparent in the long term). Both issues can pose serious problems when attempting to gauge the benefits of P4P since it is plausible that the early stages of new schemes could disrupt services and even harm performance in the short term.

- *Data of interest to evaluators might be difficult to access or non-existent.* Many P4P evaluations have relied on existing performance monitoring systems. However, evaluating the success of a P4P scheme goes beyond the assessment of incentivised performance measures and encompasses other fundamental aspects such as unintended consequences (including impacts on non-incentivised areas), or the impact on inequality in access to or outcomes of care, or changes in providers' behaviour. Evaluating the impact of P4P on these aspects typically requires having access to other types of data than those available in routine monitoring systems. For example, it may be relevant to an evaluation to obtain information on outcomes beyond the scope of those incentivized by the P4P scheme and thus not routinely recorded (e.g. a scheme may be focused on raising the performance of the weakest performers, but it may be relevant to comprehensive evaluation to assess the performance of those who were initially the strongest performers and who were not eligible to receive any additional incentives.

- *Other interventions can be implemented alongside P4P, which make identification of the impact of P4P alone impossible*. This is particularly problematic in health care where public reporting of information on quality of care provided (by medical practices or hospitals) has often been introduced alongside or as a by-product of a P4P innovation. As a result, it is not known how much of any performance change is likely to relate to the offer of performance-related payments and how much to public reporting.

- *Schemes may be modified, sometimes in unexpected and poorly understood ways, in response to local population and contextual factors, thereby making it more difficult to identify precisely which P4P design or intervention is being evaluated in a particular study*. This risk appears to have risen in the last few years in England with the emphasis on permitting and encouraging localism ('local solutions to local problems') even in the context of national 'pilots' and 'demonstrations'.

Most of the challenges highlighted above are compounded in the case of retrospective P4P evaluations, which are limited to secondary analysis of existing datasets. Not only do these studies suffer from methodological limitations of non-experimental evaluations, they rarely have the (level of detailed) information needed to answer refined questions. Some of these issues might be better addressed if evaluators were more routinely involved at an early stage of the design and implementation of P4P schemes.

What are the gaps in the evidence on P4P?

There are at least four fundamental issues in designing P4P schemes as we saw earlier – what to reward, who to reward, how to reward and how to manage perverse consequences created by the P4P incentives – yet major gaps in knowledge in relation to each remain, particularly in the health care field. This strongly suggests a number of important areas for future evaluative research:

- The long-term effects of P4P schemes, and particularly whether they can sustain behaviour (and organisational) change among providers, particularly whether they can maintain performance once performance has improved, and at which point (positive) effects appear after schemes begin (and disappear after incentives are removed or modified [62]).

- The effects of different combinations of features of P4P schemes studied simultaneously in the same or similar contexts and perhaps systematically varied over time in a manner analogous to adaptive clinical trials (e.g. to identify 'dose-response' relationships as different percentages of funding are put at risk). This could include multi-arm studies that attempt to disentangle the relative contribution of performance incentives themselves from features commonly associated with schemes, in particular, competition between providers, public reporting of results and intensive external monitoring of activities affecting activities such as client selection and reporting of outcomes. One aim of such research would be to tease out how P4P schemes work, as well as whether and to what extent. Another aim is to begin to untangle the reasons why the performance of evaluated P4P schemes appears to be so heterogeneous.

- The influence of different contexts and starting points on the performance of schemes.

- There is limited evidence on the unintended positive and negative effects of P4P schemes, partly because evaluations often rely on data produced by P4P schemes themselves (e.g. for monitoring), which, by definition, do not include the impact on non-incentivised activities. There are also potentially important effects such as 'cherry picking' of easier clients to improve outcomes and increase profits.

- To date there is very limited evidence on the effects of P4P mechanisms on intrinsic motivation which is known to be prominent among health care providers.

There is some evidence that suggests that health care providers may find that P4P undermines their motivation to act as autonomous professionals [63].

- Most of the evidence on the effects of P4P schemes in health care comes from non-experimental studies and with volunteers, making it difficult to isolate the causal effect of the schemes themselves from other factors. Greater efforts need to be made to encourage use of quasi-experimental and randomised evaluation designs involving 'cases' and 'controls' selected at random from among volunteer organisations, teams and individuals.

- The question of whether P4P schemes that target processes plausibly related to final outcomes rather than final outcomes themselves are more successful in achieving better health outcomes is still unanswered. Most evaluations focus on the effectiveness of the P4P scheme in improving the indicators rewarded by the P4P programme. It is less common to explore whether other performance measures, in particular, less proximate (and harder to measure) health outcomes, also improve.

- According to a systematic review published in 2012 [64], the evidence on the cost-effectiveness of P4P in health care is *"scarce and inconclusive"*. The authors could only identify three relatively thorough economic evaluations of P4P, and several other studies with methodological flaws, such as including only a limited range of costs and outcomes. This is important given the reports from many P4P schemes drawing attention to the complexity and unexpected costs of setting up and running schemes.

- Many P4P schemes in health care aim to incentivize teams (e.g. GP practices, specialty teams in hospitals). In order to achieve P4P rewards, these teams or organisations, probably have to make some changes to their care processes. Yet there is little knowledge about how P4P schemes targeting organisations or teams are successful, when they are. One area for future research is to understand how organisations and teams react to the implementation of P4P and how they are able to achieve the targets set for them, or not. Studies should also begin to look at the (cost) effectiveness of schemes at different levels in systems, including their effects on individual members of provider staff and clients, identifying the nature of the links between incentives at different levels on cost-effectiveness.

Conclusions

There has been a growing interest in, and use of, "pay-for-performance" in public services including health care, in the form of financial incentives (disincentives) rewarding health care providers for (not) meeting certain performance targets related to quality, efficiency and other objectives. So far, the evidence on the use of such mechanisms in health care shows mixed results and the jury is still out on the benefits of such schemes especially in relation to their costs.

Three factors are likely to contribute to the difficulty in reaching an easy conclusion on this issue.

First, the health care sector appears to be a particularly complex setting in which to introduce P4P. As with many other public services, health care has a number of distinctive features that make the design of effective financial incentives less obvious than in a typical private and/or commercial setting. Health care providers frequently bring strong intrinsic motivation to their work, striving to provide good quality services to patients in ways that are not straightforwardly related to incentives at either the individual or organisational level. In such circumstances, if incentives are perceived as controlling, they risk being less effective. Providers face a series of objectives that can be sometimes contradictory (e.g. quality of care versus efficiency of resource use) and they are accountable to varying degrees to different 'principals' (i.e. patients, their managers, the purchaser). There is still much debate as to whether provider incentives through P4P schemes are the best way to improve health outcomes. When outcomes are determined not only by providers' decisions but also by health systems' structural characteristics (e.g. poor accountability and governance) or patients' behaviours or lifestyles, incentives targeting providers may not be the most relevant interventions to improve health.

Second, the potential complexity of the technical details of P4P schemes challenges the idea sometimes conveyed that P4P is a generic intervention. There are at least three main dimensions along which financial incentives can vary for health care providers: deciding who the recipient of the bonus should be, which activities and related outcomes are to be rewarded by the scheme, and how the payment mechanism is to be structured (e.g. at which level of achievement payments are triggered). Each of these decisions will generate its own set of potential effects (sometimes contradictory), which together will shape the nature and strength of the incentives created for health care providers to perform in the desired direction and reach pre-set targets. None of these technical features is trivial, so it is possible to build many different interventions on the basis of the shared idea of paying for performance rather than paying simply for service delivery or capacity.

Third, contextual obstacles or facilitating factors at the implementation phase add to the technical complexity of P4P schemes to make each one virtually unique. The limited evidence about challenges to implementing financial incentives points to two potential issues. On the one hand, the signals theoretically created by well-designed schemes can be blurred if the conditions for a smooth implementation are not met. For example, lack of understanding on the part of front-line providers or absence of appropriate and timely feedback (through adequate data collection

systems) has the potential to render financial incentives ineffective. On the other hand, strong involvement of local actors without sufficient central direction and technical input risks producing a multitude of locally tailored schemes that stand little chance of being effective (e.g. because they are not designed on the basis of the experience of previous P4P schemes or the actions they incentivize are unlikely to produce the positive effects sought). Assessing the effects of a particular intervention will be virtually impossible in such cases as each local version will be unique, with no real 'control' to compare it with. A good balance between implementers and designers of schemes is necessary to make sure that P4P schemes are well understood (including their potential to generate unintended consequences), well designed, and well implemented.

References

1. Burgess, S. and M. Ratto, *The Role of Incentives in the Public Sector: Issues and Evidence*. Oxford Review of Economic Policy, 2003. **19**(2 (16)): p.285-300.
2. Francois, P. and M. Vlassopoulos, *Pro-social Motivation and the Delivery of Social Services*. CESifo Economic Studies, 2008. **54**(1): p.22-54.
3. Van Herck, P., et al., *Systematic review: Effects, design choices, and context of pay-for-performance in health care*. BMC Health Services Research, 2010. **10**(1): p.247.
4. Lazear, E., *Performance pay and productivity, in NBER working paper 56721996*. National Bureau of Economic Research: Cambridge MA. p.34.
5. Lazear, E., *The power of incentives*. American Economic Review, 2000. **90**(2): p.410-414.
6. Abel, P. and A. Esmail, *Performance pay remuneration for consultants in the NHS: is the current system fair and fit for purpose?* J R Soc Med, 2006. **99**(10): p.487-93.
7. Kreps, D.M., *Intrinsic motivation and extrinsic incentives*. American Economic Review 1997. 87(2): p.359-364.
8. Le Grand, J., *Motivation, Agency and Public Policy: Of Knights and Knaves*, Pawns and Queens. 2003, Oxford: Oxford University Press.
9. Frey, B. and R. Jegen, *Motivation crowding theory*. Journal of Economic Surveys, 2001. **15**(5): p.589-611.
10. Basinga, P., et al., *Effect on maternal and child health services in Rwanda of payment to primary health-care providers for performance: an impact evaluation*. The Lancet. **377**(9775): p.1421-1428.
11. OECD, Improving value for money in health by paying for performance, in Value for money in health spending2010, OECD publishing: Paris.
12. Sutton, M., et al., *Reduced mortality with hospital pay for performance in England*. N Engl J Med., 2012. **367**(19): p.1821-8. doi: 10.1056/NEJMsa1114951.
13. Woolhandler, S., D. Ariely, and D.U. Himmelstein, *Why pay for performance may be incompatible with quality improvement*. BMJ., 2012. **345:e5015**.(doi): p.10.1136/bmj.e5015.
14. Mannion, R. and H.T. Davies, *Payment for performance in health care*. BMJ, 2008. **336**(7639): p.306-8. doi: 10.1136/bmj.39463.454815.94.
15. Glasziou, P.P., et al., *When financial incentives do more good than harm: a checklist*. BMJ, 2012. **345:e5047**.(doi): p.10.1136/bmj.e5047.
16. Werner, R.M., et al., *The effect of pay-for-performance in hospitals: lessons for quality improvement*. Health Aff (Millwood). 2011. **30**(4): p.690-8. doi: 10.1377/hlthaff.2010.1277.
17. Fleetcroft, R. and R. Cookson, *Do the incentive payments in the new NHS contract for primary care reflect likely population health gains?* Journal of Health Services Research & Policy, 2006. **11**(1): p.27-31.

18. Campbell, S.M., et al., *Effects of pay for performance on the quality of primary care in England*. N Engl J Med, 2009. **361**(4): p.368-78.

19. Pauly, M. and A. Swanson, *Social Impact Bonds in Nonprofit Health Care: New Product or New Package?*, 2013.

20. Liebman, J.B., *Social Impact Bonds: A Promising New Financing Model to Accelerate Social Innovation and Improve Government Performance*, 2011: Center for American Progress.

21. Heath, C., *On the Social Psychology of Agency Relationships: Lay Theories of Motivation Overemphasize Extrinsic Incentives*. Organ Behav Hum Decis Process, 1999. **78**(1): p.25-62.

22. Rosenthal, M.B. and R.A. Dudley, *Pay-for-performance: will the latest payment trend improve care?* JAMA, 2007. **297**(7): p.740-4.

23. Kahneman, D. and A. Tversky, *Prospect theory: an analysis of decision making under risk*. Econometrica, 1979. **42**: p.263-291.

24. Maynard, A., *The powers and pitfalls of payment for performance*. Health Econ., 2012. **21**(1): p.3-12. doi: 10.1002/hec.1810.

25. Jacob, B.A. and S.D. Levitt, *Rotten Apples: An Investigation of The Prevalence And Predictors of Teacher Cheating*. Quarterly Journal of Economics, 2003. **118**(3): p.843-878.

26. Holstrom, B. and P. Milgrom, *Multitask principal-agent analyses: incentives contracts, asset ownership and job design*. Journal of Law, Economics and Organization, 1991.**7**: p.24-52.

27. Eijkenaar, F., *Pay for Performance in Health Care: An International Overview of Initiatives*. Medical Care Research and Review, 2012.

28. Infante, A., M. Meit, and E. Hargrave, *Medicare Physician Quality Reporting Initiative: Implications for Rural Physicians (Final Report) 2010*, NORC Walsh Center for Rural Health Analysis.

29. Hillman, A.L., et al., *Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care*. Am J Public Health., 1998. **88**(11): p.1699-701.

30. Greenhalgh, T., et al., *Diffusion of innovations in service organizations: systematic review and recommendations*. Milbank Q, 2004. **82**(4): p.581-629.

31. Christianson, J.B., D.J. Knutson, and R.S. Mazze, *Physician pay-for-performance*. Implementation and research issues. J Gen Intern Med, 2006. **21 Suppl 2**: p.S9-S13.

32. McDonald, R., et al., *Evaluation of the Commissioning for Quality and Innovation Framework*, 2013, University of Nottingham: Nottingham, UK.

33. McDonald, R., et al., *Changes to financial incentives in English dentistry 2006–2009: a qualitative study*. Community Dentistry and Oral Epidemiology, 2012. **40**(5): p.468-473.

34. Maisey, S., et al., *Effects of payment for performance in primary care: qualitative interview study*. Journal of Health Services Research & Policy, 2008. **13**(3): p.133-139.

35. Kirschner, K., et al., *Design choices made by target users for a pay-for-performance program in primary care: an action research approach*. BMC Family Practice, 2012. **13**(1): p.25.

36. Prentice, G., A.W. Burgess, and C. Propper, *Performance pay in the public sector: A review of the issues and evidence*, 2007, Office of Manpower Economics.

37. Muralidharan, K. and V. Sundararaman, *Teacher performance pay: experimental evidence from India, in Working paper no. 153232009*, National Bureau of Economic Research, Cambridge, MA.

38. Glewwe, P., N. Ilias, and M. Kremer, *Teacher Incentives*. American Economic Journal: Applied Economics, 2010. **2**(3): p.205-27.

39. Lavy, V., *Evaluating the effect of teachers' group performance incentives on pupil achievement*. The Journal of Political Econo, 2002. **110**(6): p.1286-1317.

40. Lavy, V., *Performance pay and teachers' effort, productivity and grading ethics*. American Economic Review, 2009. **99**(5): p.1979-2011.

41. Springer, M.G., et al., *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*, 2010, National Center on Performance Incentives at Vanderbilt University: Nashville, TN.

42. Springer, M.G. and M. Winters, *New York City's School-Wide Bonus Pay Program: Early Evidence from a Randomized Trial*, 2009, National Center on Performance Incentives at Vanderbilt University: Nashville, TN.

43. Eberts, R., K. Hollenbeck, and J. Stone, *Teacher Performance Incentives and Student Outcomes*. Journal of Human Resources, 2002. **37**(4): p.913-927.

44. Figlio, D.N. and L.W. Kenny, *Individual teacher incentives and student performance*. Journal of Public Economics, 2007. **91**(5-6): p.901-914.

45. Dee, T.S. and B.J. Keys, *Does Merit Pay Reward Good Teachers? Evidence From a Randomized Experiment*. Journal of Policy Analysis and Management, 2004. **23**(3): p.471-488.

46. Atkinson, A., et al., *Evaluating the impact of performance-related pay for teachers in England*. Labour Economics, 2009. **16**(3): p.251-261.

47. Burgess, S., et al., *Incentives in the Public Sector: Evidence from a Government Agency, in Working Paper 11/2652011*.

48. Burgess, S., et al., *Smarter Task Assignment or Greater Effort: The Impact of Incentives on Team Performance**. The Economic Journal, 2010. **120**(547): p.968-989.

49. Greene, S. and D. Nash, *Pay for performance: an overview of the literature*. Am J Med Qual, 2009. **24**(2): p.140-163.

50. Petersen, L.A., et al., *Does Pay-for-Performance Improve the Quality of Health Care?* Ann Intern Med, 2006. **145**(4): p.265-272.

51. Town, et al., *Economic incentives and physicians' delivery of preventive care*. American Journal of Preventive Medicine, 2005. **28**(2): p.234-240.

52. Christianson, J., S. Leatherman, and K. Sutherland, F*inancial Incentives, Healthcare Providers and Quality Improvements: A Review of the Evidence*, 2007, The Health Foundation: London.

53. Gillam, S.J., A.N. Siriwardena, and N. Steel, *Pay-for-Performance in the United Kingdom: Impact of the Quality and Outcomes Framework – A Systematic Review*. The Annals of Family Medicine, 2012. **10**(5): p.461-468.

54. Scott, A., et al., *The effect of financial incentives on the quality of health care provided by primary care physicians*. Cochrane Database Syst Rev., 2011(9): p. CD008451. doi: 10.1002/14651858.CD008451.pub2.

55. Mehrotra, A., et al., *Pay for Performance in the Hospital Setting: What Is the State of the Evidence?* American Journal of Medical Quality, 2009. **24**(1): p.19-28.

56. Houle, S.K.D., et al., *Does Performance-Based Remuneration for Individual Health Care Practitioners Affect Patient Care? A Systematic Review*. Annals of Internal Medicine, 2012. **157**(12): p.889-899.

57. Doran, T., et al., *Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework*. BMJ, 2011. **342**.

58. Gravelle, H., M. Sutton, and A. Ma, *Doctor behaviour under a pay for performance contract: treating, cheating and case finding?* Economic Journal, 2010. **120**: p.F129-F56.

59. Roski, J., et al., *The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines*. Prev Med, 2003. **36**(3): p.291-9.

60. McDonald, R., et al., *A Qualitative and Quantitative Evaluation of the Introduction of Best Practice Tariffs*, 2012, University of Nottingham: Nottingham, UK.

61. Imbens, G.W., *Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review*. Review of Economics and Statistics, 2004. **86**(1): p.4-29.

62. Lester, H., et al., *The impact of removing financial incentives from clinical quality indicators: longitudinal analysis of four Kaiser Permanente indicators*. BMJ, 2010. **340**.

63. McDonald, R. and M. Roland, *Pay for performance in primary care in England and California: comparison of unintended consequences*. Ann Fam Med., 2009. **7**(2): p.121-7. doi: 10.1370/afm.946.

64. Emmert, M., et al., *Economic evaluation of pay-for-performance in health care: a systematic review*. Eur J Health Econ., 2012. **13**(6): p.755-67. doi: 10.1007/s10198-011-0329-8. Epub 2011 Jun 10.

PIRU POLICY INNOVATION RESEARCH UNIT

The Policy Innovation Research Unit (PIRU) brings together leading health and social care expertise to improve evidence-based policy-making and its implementation across the National Health Service, social care and public health.

We strengthen early policy development by exploiting the best routine data and by subjecting initiatives to speedy, thorough evaluation. We also help to optimise policy implementation across the Department of Health's responsibilities.

Our partners

PIRU is a novel collaboration between the London School of Hygiene & Tropical Medicine (LSHTM), the Personal Social Services Research Unit (PSSRU) at the London School of Economics and Political Science (LSE), and the Health and Care Infrastructure Research and Innovation Centre (HaCIRIC) at Imperial College London Business School plus RAND Europe and the Nuffield Trust.

The Unit is funded by the Policy Research Programme of the Department of Health.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Imperial College London BUSINESS SCHOOL

LSE PSSRU

RAND EUROPE

nuffieldtrust

Policy Innovation Research Unit
Department of Health Services Research & Policy
London School of Hygiene & Tropical Medicine
15–17 Tavistock Place, London WC1H 9SH

Tel: +44 (0)20 7927 2784
www.piru.ac.uk