Monitor
Making the health sector work for patients

Research on financial and non-financial incentives

www.gov.uk/monitor

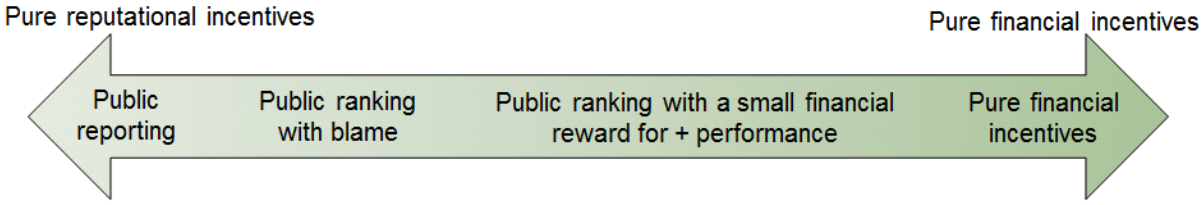
Introduction

A wide range of factors impact the behaviour of individual people and organisations ('actors') within the NHS, and in turn this behaviour affects patient outcomes. These factors can take the form of incentives that promote or discourage certain activities. Payment is one of them, but many other factors can impact or interact with payment incentives to enhance or diminish their effect. Therefore, in considering payment reform for NHS services, it is vital to recognise that the ability for the payment system to enable desired outcomes can be impacted by other types of incentives within the system, and payment incentives need to be developed within the broader context of other types of incentives.

Key findings and implications

As part of Monitor and NHS England's development work on reforming the payment system for NHS services, Monitor was keen to develop a better understanding of what mix of financial and non-financial incentives may drive the best outcomes for patients and enable a more sustainable and responsive NHS. As part of this work we looked at international examples, journal publications and sought input from NHS stakeholders (including a crowdsourcing exercise).[1]

We also commissioned research, as summarised in the paper 'Incentivising Improvements in Health Care Delivery' (conducted by Dr Adam Oliver, Reader in the Department of Social Policy at the London School of Economics). It explores four different types of incentive that sit along a scale from the purely reputational to the purely financial:



1. Pure financial incentives for predefined improvements against specific quality criteria.
2. Public ranking of performance and modest financial incentives for good relative performance.
3. Public ranking of performance with apportioned blame for poor relative performance.
4. Public reporting of general performance without financial incentives.

[1] See 'February 2014 crowdsourcing exercise on the design of the NHS payment system', available at https://www.gov.uk/government/publications/reforming-the-payment-system-for-nhs-services-supporting-the-five-year-forward-view

The paper suggests that incentives within the payment system could have an important role to play in effectively driving change in specific and well defined areas. However, these could be complemented by non-financial incentives (eg reputational) which can also be effective in motivating service delivery improvement.

Key implications of the research include:

- **Balance of different incentives:** A mix of well-designed financial and non-financial incentives are likely to be most effective. Financial incentives that offer a small financial reward (as opposed to threatening financial penalties) may best encourage innovation and organisational change within the sector.
- **Benchmarking:** Public rankings and benchmarking against other teams or organisations can be effective, but need to be managed in a way that ensures they are used constructively to promote continued learning and improvement, and do not damage morale.
- **Impact on different actors:** Incentives that are designed to operate at an organisation level must flow through to have an impact on the behaviour of the individuals who make the day–to-day decisions that ultimately determine the care that patients receive.
- **Innovation:** Incentives that create an environment of risk aversion may have an adverse impact on people innovating to improve service delivery.

Incentivising Improvements in Health Care Delivery

Adam Oliver, PhD*

Department of Social Policy, London School of Economics, Houghton Street, London WC2A 2AE, UK

Introduction

Over the past fifteen years, performance management has become an increasingly applied instrument in health care organisations in a number of countries, and is the principal method by which to incentivise changes in behaviour. In the United States (US), an early and continuing innovator in this area, the implementation of performance management grew in the late 1990s in part as a response to reports of avoidable medical mortality and practice variation, mirrored at around the same time in the UK by, for example, the public inquiry into paediatric cardiac deaths in Bristol (Department of Health, 2001). A new era, focusing more heavily on patient safety, was born, and professional and organisational accountability was highlighted.

Some felt that performance management, by using feelings of self-interest to motivate 'best practice', would drive improvements in quality. Performance management can, however, take a number of qualitatively different forms, the effectiveness of which might rely on different cognitive responses, and each of which may have different negative unintended consequences. Four of the main methods of performance management, broadly defined, are as follows:

1. Pure financial incentives for predefined improvements against specific quality criteria.

2. Public reporting of general performance without financial incentives.

3. Public ranking of performance with apportioned blame for poor relative performance.

4. Public ranking of performance and modest financial incentives for good relative performance.

Although much further work needs to be undertaken in order to conclude more definitively on which, if any, of these four types of performance management offers promise in terms of improving

* Tel.: +44-(0)20-7955-6471; *E-mail address*: a.j.oliver@lse.ac.uk

quality at reasonable cost, there is enough extant literature to enable us to make some educated speculations. Each of the four types will thus now be considered in turn.

Pure financial incentives

Pure financial incentives are better known as pay for performance (P4P). In theory, P4P is a simple instrument that appeals to a straightforward human response, informed by the relative price mechanism of standard economic theory: i.e. if you pay someone to do a particular thing, they are more likely to do it. Unfortunately, the effectiveness of P4P in applied settings, let alone its cost-effectiveness, is less clear cut – Tanenbaum (2009) cites evidence that gives a mixed picture of the effectiveness of P4P. For instance, Rosenthal and Frank (2006), in a review of applications both inside and outside health care, found little evidence of positive effects on quality. Moreover, and more specifically, although Levin-Scherz *et al.* (2006) found that, under a P4P contract, Partners HealthCare System in Boston significantly improved on its performance indicators for diabetes care compared with Partners and non-Partners plans in other locations, these improvements were not replicated for performance in pediatric asthma care.

Many P4P applications might, however, have been insufficiently powered to show much effect. If people are paid enough to do something, we can probably have quite a reasonable degree of confidence that they will do it, a powerful illustration of which has been observed in NHS primary care. In 2004, the government introduced an element of P4P into the contract for paying general practitioners (GPs), by making part of their remuneration dependent on performance against (initially) 146 indicators of clinical quality, practice organization, and patient experience – i.e. the quality and outcomes framework, or QOF. The new P4P incentives were comprised of mostly additional payments, worth up to £1 billion per year in total, 20 percent of the total GP budget at that time (Roland, 2004). Thus, presumably in order to get the acquiescence of the GPs, there were no immediate losers (99.6 percent of GPs signed up for the new contract, even though participation was voluntary (Campbell *et al.*, 2009)).

Doran *et al.* (2006) observed that during the first year of this P4P mechanism, across the quality indicators, an average of 83.4 percent of patients were assessed against them. GP practices earned an average of £76,200 from the performance mechanism, which greatly exceeded what the government had anticipated. Prior to the 2004 contract, each GP typically earned £70,000 – £75,000 per annum; after the introduction of P4P, their average income rose by £23,000. The GPs clearly

responded to the incentives, which, if the chosen quality indicators genuinely improved the quality of care, and there was not substantial gaming activity by GPs, presumably improved primary health care delivery, at least over the period studied.

Although Doran *et al.* (2006) did not find significant evidence of GPs gaming the system by inappropriately excluding patients who have missed the targets, other forms of gaming cannot of course be ruled out. For instance, pay for performance may still encourage a focus on relatively healthy patients – e.g. if one of the targets is to control cholesterol below a particular level for a certain proportion of patients, one might concentrate on reducing cholesterol in those who are just above the threshold (with a possible increased readiness to prescribe cost-ineffective care), even though population health might be better served by trying to reduce cholesterol for those with levels far above the threshold. Similarly, performance incentives may take time away from sicker or less compliant patients, whose indicators are harder to improve. It is possible that more blatant forms of cheating will also occur, such as doctors recording lower blood pressure measurements than their patients actually have, perhaps necessitating a process of aggressive monitoring and inspection, which inevitably adds to costs.

Moreover, regarding the apparent effectiveness of the GP contract, Campbell *et al.* (2007) struck a note of caution by pointing out that a range of initiatives that had been implemented before the introduction of P4P, including national standard setting for the treatment of major chronic diseases, were already contributing to improvements in process quality. They found that, for asthma, diabetes and coronary heart disease, care had improved significantly better than the longer-term trend for the former two illnesses between 2003 and 2005, but not for the latter. Campbell *et al.* (2009) reported a follow-up study, and found that by 2007 the rate of improvement had slowed down for all three conditions, to the point where the improvements were increasing at only the pre-2004 rate. The authors suggest several possible reasons for why the improvement in performance slowed down, including the fact that near-maximal performance scores had already been achieved against at least some of the criteria, and that the structure of the incentive mechanism did not reward improvements that exceeded the initial targets. This demonstrates that P4P mechanisms have to be cleverly designed in order to try to secure a sustained effect, but also places a cautionary mark against relying on evidence of short term effect as support for advocating strongly for any behavioural change initiative.

On balance, it appears that P4P can be effective in motivating people to perform incentivised actions, if the incentive is meaningful to them and outweighs the inconveniences that a change in

behaviour entails, at least for as long as the incentive remains meaningful to them and there is sufficient scope for them to change their behaviour (e.g. if they are doing the best that they realistically can against a performance indicator, then they are unlikely to be able to improve further). The design of the incentive is therefore a key consideration. Moreover, it may be the case that P4P is potentially most effective when targeted specifically at individuals (e.g. GPs) in relation to tightly specified discrete actions, rather than at the level of general organisational-level change. For pure financial incentives to be meaningful, however, comes with considerable cost implications, which places a question mark against the cost-effectiveness of this instrument. Even if they were found to represent good value for money, their impact on the health care budget remains a highly relevant consideration in a cost-constrained environment. We might therefore find it fruitful to turn to other forms of performance management where effectiveness might not depend on such a substantial monetary input.

Reporting general performance

Simply requiring hospitals to report their general performance to some higher organisational body, or face a financial penalty for not doing so, might be expected to motivate performance improvements, because the hospital managers may perceive that they are being in some sense monitored, and will not want to give the impression that their organisation is performing poorly. Since the mechanism does not reward or punish actual performance – rather, just the reporting of it – the expectation of performance improvement does not seem to be informed by standard economic theory. It may loosely be informed by a psychological finding known as priming. A study undertaken by Bateson *et al.* (2006) helps to illustrate this phenomenon. In their study, office workers were allowed to help themselves to tea or coffee in a shared kitchen, on the understanding that the workers would make voluntary contributions to pay for further supplies. For a period of ten weeks, a poster was displayed next to the suggested price list, with the picture alternating each week between an image of flowers, and an image of a pair of eyes. The contributions were three times higher in the 'eye weeks' than in the 'flower weeks', indicating that the symbolic reminder of being watched was sufficient to motivate people to contribute more money. Importantly, the priming literature has generally not reported evidence of sustained effectiveness, but we can speculate that if requiring hospitals (or other health care organisations) to report their performance gives the managers a feeling of being watched, then this might incentivise them to strive harder to improve their organisation's performance in those particular activities.

Although relatively inexpensive to implement, however, the evidence of performance improvement consequent on motivating organisations merely to report is not auspicious. For illustrative purposes, consider Medicare, the US publicly-financed health care program for the over sixty-fives. Medicare has collected hospital performance data since the early 2000s, and since 2003 has financially penalized hospitals if they fail to provide information on particular measures of clinical quality. By 2004, 98 percent of hospitals participated in the program (ACHP, 2005). One can to some extent attempt to observe whether this increased reporting has improved performance by looking at the Healthcare Effectiveness Data and Information Set (HEDIS) data reported by the National Committee for Quality Assurance (NCQA). HEDIS measures the performance of managed care plans (to which hospitals are aligned) against a number of performance criteria of the type collected by Medicare – Table 1 summarizes the performance of the Medicare plans against a selection of these criteria in 2003 and 2007. The table shows that performance deteriorated between 2003 and 2007 on some indicators and improved on others, but, with the exception of annual renal tests for diabetics, there is not overwhelming evidence that performance had, in general, improved.

Table 1

Percentage of Eligible Patients Experiencing Select Indicators of Quality

	2003	2007
Mammography	74.0	67.3
Colorectal Cancer Screening	49.5	50.4
Influenza Vaccination*	74.4	68.6
Annual HbA1c Test for Diabetes	87.9	88.1
Poor HbA1c Control for Diabetes**	23.4	29.0
Semiannual Lipid Screening	91.1	85.7
LDL Cholesterol < 100mg/dL	41.9	46.8
Annual Eye Test for Diabetics	64.9	62.7
Annual Renal Test for Diabetics	53.6	85.7
Blood Pressure ≤140/90	61.4	57.7
30-Day Follow-up of Mental Patients	60.3	54.4
6 Months of ß Blockers following MI***	61.3	75.5

Source: NCQA (2008)

*For patients aged 65 years and over

**Lower is better

***Myocardial infarction

Admittedly, this is an imperfect test of the effect of public reporting. For instance, other concomitant Medicare policy initiatives may well have confounded the results, swallowing the effect of reporting, and, moreover, in the counterfactual, without reporting, it is possible that Medicare could have fared worse than it did. Having said this, it is noteworthy that others have similarly stated that the mere publication of reported performance data often has little effect (e.g. see Besley *et al.* (2009), Marshall *et al.* (2003) and Fung *et al.* (2008)). This particular form of reporting may thus lack sufficient motivational power, although this is not to suggest that the priming phenomenon ought to be dismissed outright; indeed, priming health care professionals to undertake discrete beneficial acts, such as more regular hand washing, might prove to be a useful line of inquiry, and health care organisations should perhaps therefore be encouraged to experiment with initiatives informed by priming within their work environments. In terms of performance management, however, the mere reporting of performance with no incentives to improve performance appears to be associated with low costs but poor effectiveness. To improve motivational power, the data needs to be presented in such a way as to make it easier for the public and for the health care professionals themselves to discern how health care organisations or even individual professionals are performing relative to their peers, and to ground the policy approach more firmly in behavioural economic theory. Fortunately, there are examples of such initiatives that have demonstrated reasonable effect.

Public ranking with apportioned blame

Publishing the relative performance of health care organisations and/or professionals in an open and easily understood way introduces the concepts of reference points and loss aversion, which are among the strongest findings in the field of behavioural economics (see, for instance, Kahneman and Tversky (1984) and Tversky and Kahneman (1991)). The finding is that when people make choices, they often anchor on something that is salient to them – that is, their reference point – and act as if the avoidance of losses around that reference point is a more powerful motivating force than the possibility of experiencing a gain above the reference point. Hence, they are particularly averse to the possibility of experiencing a loss (hence the term, loss aversion): far more so than can be explained by standard economic theory. Specifically, in contexts that involve money outcomes (which is the domain over which the main findings of empirical behavioural economics were original uncovered), the disutility that individuals seemingly suffer from losses is approximately

twice as great in magnitude as the utility that they enjoy from gains of the same absolute size. Knowing that people are strongly averse to losses can be a powerful piece of information when designing public policy initiatives.

Therefore, it can reasonably be hypothesised that publishing the performance of health care entities in the form of a league table – with better relative performance winning a higher place in the table – may encourage managers and doctors to adjudge their performance against what they perceive to be their reference point (e.g. the performance of an alternative local hospital, a peer, a reasonable position set by government etc.). If current performance is poorer than the reference point, this may serve as a powerful motivator to try to improve performance in time for the publication of the next league table. Those who are performing well might be similarly motivated to retain, at the very least, their position – hence, to not lose status (admittedly, if a person or organisation routinely expects that they will perform poorly on the league table, their reference point might be low and the motivating power of loss aversion could be weak).

Some have looked at whether supplementing the league or performance table with some form of blame or punishment for poor relative performance might strengthen the instrument still further, in essence, one might hypothesise, by further strengthening the aversion to losses. We can denote this naming and shaming concept as blame for performance, or B4P. A good example of this type of incentive mechanism was used in in the NHS hospital sector in the early to mid-2000s. Specifically, in 2001 the government introduced a performance framework called the hospital star rating system, whereby NHS hospitals in England were assessed annually on a number of indicators, including targets against waiting times, cleanliness, treatment-specific data and financial management (the star rating system was replaced by another system of reporting hospital performance against waiting time targets from 2006 to 2010 (called the Annual Health Check), although the Labour Government emphasized a new policy of patient choice during that latter period). After assessment, hospitals were each awarded from zero to three stars, with more stars indicating better performance, and, perhaps, at least two stars serving as a reference point for many providers. The star ratings were publicised in national and local media, and for very poor performance hospital management teams could be dismissed (for very good performance, hospitals could earn greater autonomy from central government, and therefore the instrument did not rely on punishment exclusively as a motivator for behaviour change).

The threat of dismissal and the fact that relative performance was fairly widely publicized clearly demonstrate that the star rating system was a ranking exercise that apportioned blame. Other NHS

policies, such as the patient choice policy, may confound the effectiveness of the star rating system to some extent (although the patient choice policy was not introduced until January 2006), but several authors have lauded the effectiveness of star rating, particularly with respect to reducing waiting times (e.g. Besley *et al.*, 2009; Bevan and Hood, 2006; Propper *et al.*, 2008). Mays *et al.* (2011) concluded from a review of these various reforms that accountability against targets, at least in the quite narrow domains over which targets were set, appeared to be more effective than choice and competition. Moreover, using the star rating system as a natural experiment, Bevan and Fasolo (2013) note that Wales, Scotland and Northern Ireland, with similar increases in health care spending, did not institute the same magnitude of star rating governance system of targets and reputation, and, as a consequence, did not achieve the reductions in waiting times observed in England. Initially, the principal waiting time target in the star rating system was that no patient should have to wait more than twelve months for elective hospital care. Subsequently, the target was reduced to six months and then eighteen weeks (which, by the end of the decade, had been almost met). Table 2 illustrates the striking reduction in waiting times – arguably the biggest NHS success of the last two decades – since the introduction of the star rating system.

Table 2

Trends in Waiting Times in England

	Months Waiting (% of total)		
Year	< 3	< 6	< 12
March 2000	51	74	95
March 2004	54	81	99.9
March 2008	92	100	
March 2012	97	100	

Source: Department of Health (various years)

Although it is fairly uncontroversial to conclude that the behavioural motivators instituted by the star rating system contributed substantively to the fall in waiting times, the unprecedented increase in NHS expenditures during the decade beginning in 2000 facilitated an increase in the system's capacity that was a necessary requirement for the observed decrease in waits (specifically, between 1997 and 2011, public health care expenditure increased from £44 billion to £118.3 billion, or 7.3% per annum, while average annual general inflation over the period was only 2.4% (ONS, 2013)). Whether a similar mechanism might work to the same extent in a resource-constrained environment

is, at least, questionable. Moreover, a sustained use of publicising relative performance alongside naming and shaming when there are resource constraints may negatively impact on morale (which one may expect from a policy that has been termed colloquially as 'targets and terror'), undermining the ability and willingness of NHS staff to identify with the system.

The importance of recognising people's ability to identify with the organisation in which they work, and the effect that this may have on their performance, has been a key consideration in behavioural economics over the last ten years, in particular in the work of Akerlof and Kranton (2010). They posit that people experience positive utility from working for an organization with which they identify and negative utility if they perceive themselves to be outsiders. Identity utility is thus the gain we feel when the actions and ethos of people and things around us (our peer group, our workplace, etc.) conform to our norms and ideals; identity disutility is the converse. Although such a general concept may be familiar in other branches of social science, the utility experienced from feeling that one belongs is not generally incorporated formally into standard economic theory. Akerlof and Kranton apply their framework to firms, the military, the education sector, gender in the labour market and in the home, and race and poverty. If we take the organisational behaviour of firms, for example, according to Akerlof and Kranton, good management involves creating workers who are motivated insiders who identify with the goals of the firm, rather than alienated outsiders. Aligning the objectives of managers (or policy makers) and workers is the goal of a strategy called 'management by objective', which works by changing the self- motivation of workers. The intention is that after a while workers will want to achieve the objectives of the firm, regardless of additional personal financial rewards, because they will identify with those objectives. Akerlof and Kranton did not consider health care, but it is quite plausible that policies, such as public reporting with naming and shaming, whilst well-grounded positively in some aspects of behavioural theory, may, according to other theoretical behavioural postulates, pose the potential to damage an organisation (e.g. the NHS) in the long run. In short, it is important to recognise that behavioural economics may sometimes predict policy effects that are socially beneficial or harmful, depending on which behavioural concept dominates. Therefore, if one wants to try to incentivise with league tables, it may be prudent to focus on reward rather than punishment.

Public ranking with modest financial rewards

When faced with a performance league table where the focus is on rewards for good relative performance, behavioural economic theory might suggest that managers would still choose a

reference point and perceive performance as good or bad – i.e. a gain or a loss – relative to that reference point. Therefore, loss aversion will remain as a motivator, but the fear of punishment and the consequent harms to morale if performance is poor in relative terms is removed, or at least ameliorated (of course, if an individual or organisation is performing poorly in absolute terms, it is down to the governing authority to attempt to find out why this is the case, and, if necessary, to remove those responsible). In 2003, the Centers for Medicare and Medicaid Services (CMS), which administers both Medicare and Medicaid – the publicly financed health care system that covers many of the US's indigent – initiated an experiment that combines league table and financial incentives. The experiment, called the Hospital Quality Incentive Demonstration (HQID) project, lasted for three years, and was conducted in cooperation with Premier Inc., a non-profit hospital alliance. There were 270 hospitals of varying size in the project, located across thirty-six states, which were assessed on more than thirty performance indicators relating to acute myocardial infarction, coronary artery bypass grafting, heart failure, pneumonia, and hip and knee replacement. The hospitals were ranked annually according to their performance and those hospitals in the top decile received a 2 percent bonus on their standard DRG payments, while those in the next decile received a 1 percent bonus. Initially, therefore, the project focused on combining league tables with rewards. Admittedly, in HQID's final year, hospitals in the bottom two deciles suffered similar payment decrements, indicating that the project eventually involved a mix of positive and negative financial incentives. As noted above, whilst one may expect payment decrements to be quite motivating due possibly in part to the powerful force of loss aversion, they are punishments, and may have a net demotivating effect in the long run if they harm morale. Consequently, financially penalising poor relative performance (which may not even be poor in absolute terms) may make that performance deteriorate still further, particularly in health care organisations where the relatively poor performance is caused by extant straightened financial circumstances, and/or poor health and income profiles of the communities they serve.

Leaving aside these concerns with instituting financial penalties, Tanenbaum (2009) notes that the overall performance improvement for the project's first year was 6.6 percent. The magnitude of performance improvement required before one can conclude that an initiative has been a success is subjective, but over the three years of HQID there was an overall average improvement measuring 15.8 percent, which is not negligible (Table 3 gives a breakdown of improvement by therapeutic area). Less auspiciously, Glickman *et al.* (2007) found no significant difference between the in-hospital mortality rates for acute myocardial infarction in HQID hospitals versus controls, although this may have been because the performance indicators in this clinical area are insufficiently linked to in-hospital mortality rates, perhaps highlighting the importance of ensuring that there is a good

evidence base to show that the relatively easily influenced indicators of process quality are adequately associated with the primary outcomes of interest. A version of the HQID project has been applied in the NHS in recent years (Maynard and Bloor, 2010).

Table 3

Performance Improvement in the HQID Project*

	Inception	End of Year 3
Acute Myocardial Infarction	87.5	96.1
Coronary Artery Bypass Grafting	84.8	97.4
Heart Failure	64.5	88.7
Pneumonia	69.3	90.5
Hip and Knee Replacement	84.6	96.9

Source: U.S. Department of Health and Human Services (2008)

*The figures are the percentage of eligible patients experiencing the quality indicators

Combining league table competition with a quite modest use of financial incentives for good relative performance is also used in the US Veterans Health Administration (VHA), with apparent positive effect (Asch *et al.*, 2004; Oliver, 2007). The VHA is the largest integrated health care system in the United States, and provides public sector care for honorably discharged veterans of the US armed forces. The system is financed mostly from general taxation and is structurally very similar to the NHS. The VHA traditionally had a reputation for poor quality, which was a catalyst for a set of reforms introduced in the mid-1990s. As part of these reforms, health care managers were made accountable for performance against process quality indicators (e.g., mammography rates, LDL cholesterol < 100mg/dL etc.). Only senior managers are eligible to receive bonuses for good performance, which typically amount to up to 10 percent of their salaries, but this incentive in many cases led in turn to them putting pressure on their clinical teams to improve performance. Moreover, details of the relative performance of each facility are disseminated periodically throughout the VHA in the form of a league table, generating non-financial competition, since no hospital wants to be seen as performing worse than a local 'rival' (note that there is no competition for patients in the VHA – patients are referred to hospital by primary care gatekeepers). Other initiatives also facilitated the transformation of the VHA (Oliver, 2007), but probably principally due to the performance incentives, the VHA, within the space of 5 years, demonstrated significant improvements in process quality (see Table 4), and by 2005, with the exception of outpatient

follow-up of those admitted to hospital for a mental illness, the VHA outperformed the commercial sector, Medicare and Medicaid on all the of the quality indicators over which these systems could be compared (see Table 5).

Table 4

Summary of VHA performance at the time of the reforms vs. 5 years post reform

Type of care	*Percentage of eligible patients who experienced the quality indicator*	
	VHA (1994-95)	VHA (2000)
Preventive care		
Mammography	64	90
Influenza vaccination	28	78
Pneumococcal vaccination	27	81
Outpatient care		
For diabetes:		
Annual measurement of glycosylated hemoglobin	51	94
Annual eye examination	48	67
Semiannual lipid screening	Not reported	89
Inpatient care		
For acute myocardial infarction:		
Aspirin within 24 hrs	Not reported	93
Aspirin at discharge	89	98
Beta-blocker at discharge	70	95
ACE inhibitor if ejection fraction <40%	Not reported	90
Smoking cessation	Not reported	62
For congestive heart failure:		
Ejection fraction checked	Not reported	94

ACE inhibitor if ejection fraction <40%	Not reported	93

Source: Jha *et al.* (2003)

All of the differences 1994-95 and 2000 are significant at 0.1%.

Table 5:

Summary of VHA versus non-VHA performance in 2004-05

Type of care	*Percentage of eligible patients who experienced the quality indicator*			
	VHA (2005)	Commercial (2004)	Medicare (2004)	Medicaid (2004)
Preventive care				
Mammography	86	73	74	54
Cervical cancer screening	92	81	Not reported	65
Colorectal cancer screening	76	49	53	Not reported
Influenza vaccination*	75	Not reported	75	68
Pneumococcal vaccination	89	Not reported	Not reported	65
Outpatient care				
For diabetes:				
Annual measurement of glycosylated hemoglobin	96	87	89	76
Poor control: glycosylated hemoglobin > 9% (lower is better)	17	31	23	49
Semiannual lipid screening	95	91	94	80
Cholesterol < 100	60	40	48	31
Cholesterol < 130	82	65	71	51
Annual eye examination	79	51	67	45
Annual renal exam	66	52	59	47
For hypertension:				
BP ≤ 140/90	77	67	65	61
For mental illness:				

30 day follow-up after hospitalization	70	76	61	55
Inpatient care				
For acute myocardial infarction:				
Beta-blocker at discharge	98	96	94	85

Sources: The VHA data is reported in: VA Office of Quality and Performance (2005). The data for the commercial, Medicare and Medicaid sectors is HEDIS data.

*For patients aged 65 years and over.

Despite the promise of the admittedly small evidence base on combining modest financial incentives with league table competition in health care, a number of legitimate qualifiers can be brought forth. For instance, the positive effects observed in both the HQID project and the VHA may at least in part be the consequence of better documentation – or, colloquially, bean counting – rather than reflecting genuine improvement. Moreover, Vladeck (2004) has argued that the quality improvements are more likely to have been caused by reinforcing the norms of professional responsibility than by the inherent incentives involved, although that of course is debatable and even if true may in any case indicate that ranking with modest financial incentives are a useful way by which to reinforce those norms. Some would also rightly raise the question of whether ranking with even small incentives, even if effective, represents good value for money. However, given the paucity of cost data in this area and the uncertainty over whether process quality indicators lead to real improvements in health outcomes, not much can be said about the instrument's cost-effectiveness other than that the financial incentives used in the HQID project and the VHA, as compared with the total cost of health care delivered, were quite modest. There is thus qualified evidence of some positive effect when combining modest financial incentives with a league table ranking exercise. Ranking in and of itself may be insufficient (further research would be required to discern its independent effect), and financial incentives in and of themselves would perhaps need to be prohibitively large in the current public finances climate in order to have the desired effect (although even here, the independent effect of small financial incentives on professional and organisational behaviour merits further study).

As will be clear from what has been written up to this point, performance management initiatives have tended to focus upon motivating improvements in process quality – i.e. quite specific, discrete, easily-acted upon tasks, each of which can be undertaken by a single health care professional, such as, for example, administering vaccinations and blood sugar readings. This begs the question of whether these initiatives could be similarly deployed to improve the more complex aim of motivating a more integrated care process, which would presumably often rely on a coordinated response from multiple health care teams, who sometimes (or often?) act quite independently from one another. Even if performance indicators are restricted to the simple process indicators, performance management has been subjected to a number of criticisms that would presumably apply to an even greater extent as the indicators become more complex. For example, Roland (2004) has stated, perhaps somewhat obviously, that it is critical for the indicators to align with the policy goals one is trying to pursue (this is sometimes more difficult than it sounds – e.g. Coleman *et al.* (2007) found that pay for performance increased the likelihood of blood sugar tests but did not improve blood sugar control), but Smith (2005) has pointed out that the choice of indicator is often opportunistic and selective rather than rational, and tends to rely on existing data sources. He has thus argued that data should be collected to match the performance indicator and not vice versa. Moreover, performance management as it has tended to be applied (i.e. on a limited number of criteria, so as to avoid the possibility of target fatigue among health care professionals (Smith, 2002)), is commonly thought to be detrimental to a holistic, integrated approach to health care delivery (see, for example, Roland (2004)) because it may draw attention away from good but non-incentivised practices: i.e. good practices are plausibly crowded-out.

The evidence on this form of crowding-out is mixed. Besley *et al.* (2009) maintain that there is little evidence that the shorter waits produced by the star rating system were offset by significant detrimental effects on performance in other quality dimensions, a finding mirrored by Asch *et al.* (2004) with respect to the VHA (although in both of these studies, the authors' timeframes were quite short-term, and, moreover, they concentrated on discrete processes, and did not consider the possible negative impact on the overall integrated nature of health care delivery). However, Smith (2002), in relation to waiting-time targets that had been set before the introduction of the star rating system, reported that a preoccupation with inpatient waiting times led to widespread distortions in clinical priorities and a misrepresentation of performance. Campbell *et al.* (2009), in relation to the GP contract, noted that by 2007, the quality of those aspects of care that were not incentivized had deteriorated for asthma and heart disease care, and that the level of the continuity of care, as measured by how often patients were able to see their usual doctor, had declined.

Conceptually, assuming that we had measures that indicated a good level of integration (there seems to be no real consensus among health policy scholars of what integration actually means, let alone good data to measure this), we could replace the indicators of process quality with those measures and perhaps (in theory) expect similar improvements. But given the multifaceted coordinated response that is likely to be required to improve integration, we can have no confidence that this would occur in practice. For example, if one person or one team fails to perform satisfactorily in order to improve the overall level of integration, everyone involved in the care pathway may quickly and easily become demotivated (similar issues arise in relation to attempting to generate improvements in final health outcome rather than process, in that health outcomes are often too far away, and confounded by too many things, from what individual professionals can directly control and action – here, then, we should at least have an evidence base that demonstrates that the process quality indicator is likely to be beneficial to health outcomes in the long run). Regarding integration, possibly the best starting point would be to engage with health care professionals to attempt to find out what they believe is required to improve a coordinated response, rather than to just impose some performance indicators intended to address this issue upon them that have no bearing on their day to day professional lives. An extended consultation exercise of this kind might reveal some relatively simple actions that professionals might quite readily be able to act upon and that may be amenable to improvement via performance management initiatives, and that ultimately improve the total care experience.

Consulting widely with health care professionals when designing a performance framework serves another very important function – it helps them to identify with the instrument and may limit any possible damage to employee identity with the NHS as a whole. The VHA leadership made substantial efforts to get input from medical professionals when choosing which performance criteria to use in their performance management instrument, so as to (in a non-financial sense) 'buy' the professionals into the idea. The leadership realised that they had to make the managers and doctors feel that they were part of the plan, and even then many doctors remained opposed to performance management and left the system. In the UK, it is much more difficult for doctors to leave the NHS than it is in the multiple system US context, and thus the importance of maintaining and promoting identity is even greater, because one should want to avoid as far as possible causing greater dissatisfaction among NHS personnel.

A taxonomy of performance management initiatives

Although nothing definitive can be said about the costs and effectiveness of the different types of performance management, partly because the performance of these initiatives will be highly contextual, there is enough evidence for us to make a speculative taxonomy in this regard. This is given in Table 6.

Table 6:
Taxonomy of the costs and effectiveness of performance management

	Effectiveness	
	Reasonable	Poor
Costs		
High	Pure financial incentives (if the financial incentives are substantial – e.g. the 2004 GP contract)	
Moderate	Public ranking with modest financial rewards (e.g. the VHA) Public ranking with apportioned blame (e.g. NHS star rating)	
Low		Reporting general performance (e.g. disclosure by Medicare plans)

A cursory glance a Table 6 appears to reveal that, in an environment where there are particularly strong constraints on resources, it is advisable for policy makers to focus their attention on public ranking exercises (i.e. league table competition), either in conjunction with modest financial rewards or an element of apportioned blame (or both), if considering the application of performance managements instruments (interventions that use the notion of reference points and hence use loss aversion to motivate improvements in performance can and have been used to good effect in areas

outside of health care, such as motivating improvements in restaurant hygiene standards, encouraging healthier food purchasing patterns and more energy conscious fuel use, getting more people to pay their taxes on time, and many others; moreover, it could feasibly be used to improve performance within, as well as across, health care organisations, by instituting league table competition in the cleanliness of hospital wards, for example – consideration of the use of reference points and loss aversion is one of the strongest tools that behavioural theory has to offer to policy). However, consideration of behavioural theory might make it advisable to think twice about blaming relatively poor performers because this could serve to alienate and demotivate the very people on whom the NHS relies. Of course, poor performers in the absolute sense, such as the Mid Staffordshire NHS Foundation Trust, need to be identified, blamed and, where appropriate, punished severely, but relatively poor performance may not be bad in any absolute sense, and could be relatively poor due to circumstances beyond the control of the professionals involved (for example, due to having a disproportionately deprived case mix). It is thus incumbent on policy makers to try to better understand why the performance of those towards the bottom of the league table is relatively poor, in an attempt to identify if there is anything that can be done to help them.

The suggestion here then is that even though strengthening loss as a motivational lever is tempting, it is advisable for policy makers to use blame (or other forms of punishment, financial or otherwise) guardedly. It is at least as important to make employees (or, in other domains, students, family members, even members of society etc.) feel that that they are motivated insiders – that they share a sense of identity with the institution in question – if one wants to get the best out of them. Therefore, league table ranking, possibly with modest financial rewards, may be the most sensible application of performance management, and, plausibly, this tool could be further strengthened with a further key finding of behavioural economic theory: hyperbolic discounting or the immediacy effect, otherwise known as present bias.

Present bias is the observation that people place a heavy weight on the immediate moment, and quickly and significantly discount all future moments. The observation is generally not encapsulated in applications of standard economic theory, and can lead to a phenomenon known as dynamic inconsistency; that is, people may express a preference for a superior good that is delayed over an inferior good that is more proximate (in terms of time) if the delivery of both is promised at some future point (e.g. people might express a preference for two pieces of cake available two weeks from now over one piece of cake available one week from now), but when arriving at the time at which the seemingly inferior good can be consumed, they often switch their preference (i.e. they prefer one piece of cake if it can be consumed now over two pieces of cake available one week

from now). Thus, the apparent heavy weight that people place on the immediate moment can lead them to overemphasise the enjoyment of pleasurable experiences that occur right now (and the converse for unenjoyable experiences), and can lead to inconsistencies in preferences over time (these inconsistencies are neither predicted nor explained with use of an exponential discount rate, as is the usual practice in applied economics). Present bias is probably at least a partial explanation for many of the (in)actions that people demonstrate in their everyday lifestyle behaviours that some deem as errors (e.g. eating too much high calorific food and consuming too much alcohol and too many cigarettes due to the enjoyment that these activities offer in the immediate moment, and undertaking little exercise and saving insufficiently for retirement due to the immediate pain associated with these actions – which also in part explains procrastination bias: i.e. 'I'll put things off to tomorrow').

Knowledge of the tendency towards present bias might usefully be employed in the design of performance management initiatives, in that if one can engage those whose behaviour one is trying to influence through feelings of enjoyment, then this may increase effectiveness over and above what might otherwise be the case. Indeed, there is now a sub-field of applied behavioural science that has been dubbed fun theory that is being used, albeit in a rather ad-hoc fashion, to try to show that making things more enjoyable can improve personal lifestyle behaviours (e.g. embedding sound sensors in stairs – transforming them into 'piano stairs' – to encourage people to take the stairs rather than the escalator). There is a risk that these applications are a little gimmicky, drawing attention away from more useful but less headline grabbing initiatives, and many scholars do not approve of the word 'fun' in the academic discourse, but acknowledging the strong feelings of (dis)pleasure that people feel towards actions undertaken in the immediate moment could serve as a useful policy lever for motivating behaviour change in health care and elsewhere (much of this is not rocket science – Mary Poppins, with her spoonful of sugar, was implicitly using present bias). An additional finding in psychology is that when remembering an experience, people tend to place a heavy emphasis on the best, worst and last moments of the experience (a phenomenon known as peak-end evaluation), and underweight the duration of the event (Kahneman *et al.*, 1997). If a decision maker wants to encourage people to repeat (or abstain from) activities, knowledge of these affects are also potentially useful additions to the policy armoury.

To date, however, there seems to be little if any consideration of how phenomenon such as present bias and peak-end evaluation can be incorporated into the design of performance management initiatives, and for here this question will be left hanging, but it is perhaps noteworthy that one of the reasons for the relative success of using performance league tables might be that they, for many,

provoke some excitement associated with their natural competitive instincts. Conversely, and compounding the aforementioned concerns regarding damaging identity, it seems sensible to limit the extent to which performance management imposes stress and uncertainty on employees, because these feelings may well be magnified via present bias, and thus impose a risk of harming performance. In the context of the NHS, it may therefore be counterproductive to undermine financial security (note, the 2004 GP contract, in a time of large increases in NHS spending, fortified financial security). Thus, incentives, in straightened circumstances, might work best at the margin to encourage good professional norms and practices (with 'core' financing not subjected to the mechanism), and perhaps ought to be supplemented with league table competition since, as earlier noted, pure financial incentives may not be meaningful unless they are substantial.

References

Akerlof, G. A., and R. E. Kranton. 2010. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well- Being*. Princeton, NJ: Princeton University Press.

Asch, S. M., E. A. McGlynn, M. M. Hogan, R. A. Hayward, P. Shekelle, L. Rubenstein, J. Keesey, J. Adams, and E. A. Kerr. 2004. Comparison of Quality of Care for Patients in the Veterans Health Administration and Patients in a National Sample. *Annals of Internal Medicine* 141:938 – 945.

Alliance of Community Health Plans (ACHP). 2005. *Performance Measurement and Paying for Performance in Medicare: Health Plans, Hospitals, and Physicians*. Washington, DC: ACHP.

Bateson, M., D. Nettle and G. Roberts. 2006. Cues of Being Watched Enhance Cooperation in a Real-World Setting. *Biology Letters*, 2: 412-414.

Besley, T., G. Bevan, and K. Burchardi. 2009. *Naming and Shaming: The Impacts of Different Regimes on Hospital Waiting Times in England and Wales*. LSE Health and Social Care Discussion Paper. London: London School of Economics and Political Science.

Bevan, G., and Fasolo, B. 2013. Models of Governance of Public Services: Empirical and Behavioural Analysis of 'Econs' and 'Humans'. In: Oliver, A. (ed.). *Behavioural Public Policy.* Cambridge University Press: Cambridge.

Bevan, G., and Hood, C. 2006. What's Measured is What Matters: Targets and Gaming in the English Public Health Care System. *Public Administration* 84: 517 – 538.

Campbell, S., D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland. 2007. Quality of Primary Care in England with the Introduction of Pay for Performance. *New England Journal of Medicine* 357:181 – 190.

Campbell, S. M., D. Reeves, E. Kontopantelis, B. Sibbald, and M. Roland. 2009. Effects of Pay for Performance on the Quality of Primary Care in England. *New England Journal of Medicine* 361:368 – 378.

Coleman, K., K. L. Reiter, and D. Fulwiler. 2007. The Impact of Pay-for-Performance on Diabetes Care in a Large Network of Community Health Centers. *Journal of Health Care for the Poor and Underserved* 18:966 – 983.

Department of Health. 2001. *Learning from Bristol: Report of the Public Inquiry into Children's Heart Surgery at the Bristol Royal Infirmary (the Kennedy Report)*. London: Stationery Office.

Department of Health. Various years. *Hospital Inpatient Waiting List Statistics, England. The "Green Book."* London: Department of Health.

Doran, T., C. Fullwood, H. Gravelle, D. Reeves, E. Kontopantelis, U. Hiroeh, and M. Roland. 2006. Pay-for-Performance Programs in Family Practices in the United Kingdom. *New England Journal of Medicine* 355:375 – 384.

Fung, C. H., Lim, Y. W., Mattke, S., Damberg, C., and Shekelle, P. G.. 2008. Systematic Review: The Evidence That Publishing Patient Care Performance Data Improves Quality of Care. *Annals of Internal Medicine* 148:111 – 123.

Glickman, S. W., Ou, F. S., DeLong, E. R., Roe, M. T., Lytle, B. L., Mulgund, J., Rumsfeld, J. S., Gibler, W. B., Ohman, E. S., Schulman, K. A., and Peterson, E. D. 2007. Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction. *JAMA* 297:2373 – 2380.

Jha, A. K., Perlin, J. B., Kizer, K. W., and Dudley, R. A. 2003. Effect of the Transformation of the Veterans Affairs Health Care System on the Quality of Care. *New England Journal of Medicine* 348:2218 – 27.

Kahneman, D., and A. Tversky. 1984. Choices, Values, and Frames. *American Psychologist* 39 (4):341 – 350.

Kahneman, D., Wakker, P.P., Sarin, R., 1997. Back to Bentham? Explorations of Expected Utility. The Quarterly Journal of Economics 112:375 – 405.

Levin-Scherz, J., N. DeVita, and J. Timbie. 2006. Impact of Pay-for-Performance Contracts and Network Registry on Diabetes and Asthma HEDIS Measures in an Integrated Delivery Network. *Medical Care and Research Review* 63:14S – 28S.

Marshall, M. N., P. G. Shekelle, H. T. O. Davies, and P. C. Smith. 2003. Public Reporting on Quality in the United States and the United Kingdom. *Health Affairs* 22:134 – 148.

Maynard, A., and K. Bloor. 2010. Will Financial Incentives and Penalties Improve Hospital Care? *BMJ* 340:c88.

Mays, N., Dixon, A., and Jones, L. 2011. *Understanding New Labour's Market Reforms of the English NHS.* King's Fund: London.

National Committee for Quality Assurance (NCQA). 2008. *The State of Health Care Quality 2008.* Washington, DC: NCQA.

Oliver, A. 2007. The Veterans Health Administration: An American Success Story? *The Milbank Quarterly* 85:5 – 35.

ONS. 2013. *Expenditure on Healthcare in the UK: 2011.* Office for National Statistics: London.

Roland, M. 2004. Linking Physicians' Pay to the Quality of Care — a Major Experiment in the United Kingdom. *New England Journal of Medicine* 351:1448 – 1454.

Rosenthal, M. B., and R. G. Frank. 2006. What Is the Empirical Basis for Paying for Quality in Health Care? *Medical Care Research and Review* 63:135 – 157.

Smith, P. C. 2002. Performance Management in British Health Care: Will It Deliver? *Health Affairs* 21:103 – 114.

Smith, P. C. 2005. Performance Measurement in Health Care: History, Challenges, and Prospects. *Public Money and Management* 25:213 – 220.

Tanenbaum, S. J. 2009. Pay-for-Performance in Medicare: Evidentiary Irony and the Politics of Value. *Journal of Health Politics, Policy and Law* 34:717 – 746.

Tversky, A., and D. Kahneman. 1991. Loss Aversion in Riskless Choice: A Reference Dependent Model. *Quarterly Journal of Economics* 107:1039 – 1061.

US Department of Health and Human Services (HHS). 2008. *Premier Hospital Quality Incentive Demonstration: Rewarding Superior Quality Health Care*. Washington, DC: HHS.

VA Office of Quality and Performance. 2005. *VA's Performance Compared to Non VA.* Washington DC: Veterans Health Administration.

Vladeck, B. C. 2004. Ineffective Approach. *Health Affairs* 23:285 – 286.