HEALTH ECONOMICS
Health Econ. **21**: 3–12 (2012)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.1810

THE POWERS AND PITFALLS OF PAYMENT FOR PERFORMANCE

ALAN MAYNARD*
Department of Health Sciences, University of York, UK

1. INTRODUCTION

Throughout the world, healthcare policy makers confront common problems: expenditure inflation, inefficiency and inequity in access to care. The development of health economics during the last 20 years has produced a consensus (outside the USA) about the merits of 'single-payer' systems and the need to evaluate the cost-effectiveness of competing medical technologies. These are necessary but not sufficient conditions for expenditure control and efficient rationing (Williams, 1972; Reinhardt, 1982; Hsiao, 2011; Maynard, 1997; Culyer and Rawlins, 2004).

Recent reforms have had a modest effect on the efficiency of resource allocation in health care. Exacerbated by the global economic downturn, the desire for more radical improvements in efficiency has led to increased interest amongst policy makers in a vigorous payment-for-performance (P4P) culture based principally on the belief that financial incentives are efficient ways of mitigating variations in clinical practice and ensuring the delivery of conservative, cost-effective interventions.

The failure of public and private healthcare markets to deliver patient care efficiently, equitably and within budgets has a long history. This is reviewed in the next section and followed by a discussion of case studies of P4P, primarily in the context of healthcare provision. A selective use of this literature is used to draw out a list of central research questions to be addressed by the rapidly evolving P4P initiatives.

2. WHAT ARE THE PROBLEMS THAT PAYMENT FOR PERFORMANCE NEEDS TO ADDRESS?

Payment-for-performance programmes, such as many recent reforms in European and North American healthcare systems, aim to improve efficiency in health care. In particular, they focus on reducing variations in health care, improving processes in relation to safety and quality and increasing productivity of systems. A few of the programmes aim directly to reward healthcare outcomes. These problems are products of supplier-induced demand created by clinical autonomy and uncertainty about 'what works' in health care.

Despite the rapid growth in the healthcare industry and the evaluation of procedures used to diagnose and treat patients, the majority of healthcare interventions still lack an evidence base, and consequently, substantial *variations in health care* remain.

Figure 1 is taken from estimates about the current evidence base for all medical treatments (BMJ Evidence, 2011). It shows that less than 35% of procedures can be evidenced as beneficial or likely to be beneficial with randomised controlled trial results and systematic reviews of this work. A significant element of care has side effects, that is, may mitigate one condition and create another. Over 50% of interventions are used despite a lack of robust evidence. Investigation of these procedures is inhibited by ethical concerns (e.g. can you randomise common procedures that are believed to work?) and by the un-evidenced conservatism of medicine (we used it for years and we know it works!).

*Correspondence to: Department of Health Sciences, University of York, UK. E-mail: Alan.maynard@york.ac.uk

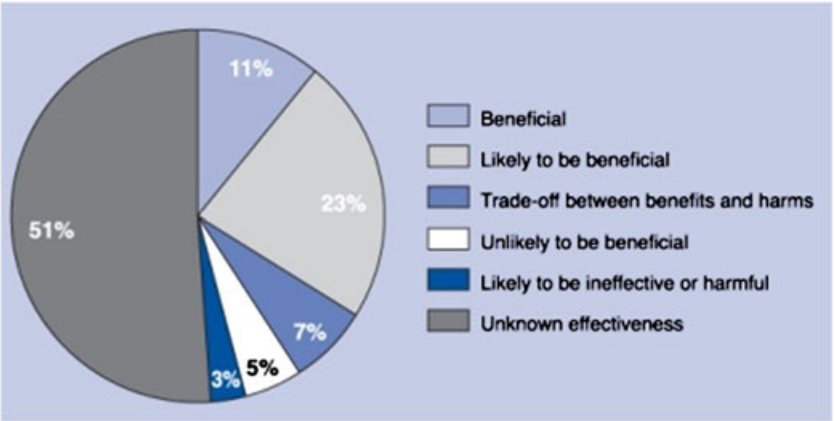

Figure 1. How much do we know? Source: BMJ Clinical Evidence 2011

If an analysis of effectiveness is focused on the proportion of medical activity that is evidence based, the picture may be more optimistic. Work from the 1990s indicates that perhaps 50% to 80% of inpatient medical care is evidence based (Ellis *et al*., 1995). However, there have been little replication of this approach. A significant gap appears to exist between clinical practice and the use of what is likely to improve patients' health.

This uncertainty about the effectiveness, let alone the cost-effectiveness of most procedures delivered everyday in healthcare systems worldwide, is one cause of the variations in clinical practice observed internationally.

Like so much in health care, this literature is many decades old, remains robustly resistant to policy reforms and fails to use insights from economics.

A curious aspect of the debate about clinical practice variations is its isolation from mainstream economics. The economics literature has shown large variations in the productivity of, for instance, the US manufacturing with firms in the best deciles producing twice as much per unit of input as firms in the worst decile. Furthermore, this literature has sought to understand better the causes of these productivity variations, for example, the role of the variable quality of management (Syverson, 2011).

The medical literature on clinical practice variations dates back nearly 75 years. Glover (1938) analysed variations in tonsillectomy rates in England, finding that the large geographical variations defied 'any explanation, save that of variations of medical opinion on the indications for the operations'. He quoted the UK Medical Research Council's conclusions that 'there is a tendency for the operation to be performed as a routine prophylactic ritual for no particular reason and with no particular result'. Tonsillectomy is a risky procedure, and Glover reported mortality with 424 schoolchildren dying following surgery in 1931–1935. Despite such risks, tonsillectomy was three times more prevalent amongst 'well to do classes', perhaps indicative of its capacity to generate personal income through fee for service payments in the days before the National Health Service (NHS).

The issue of variations in the rates of surgery for tonsils and adenoids in the UK was further analysed by Bloor and his colleagues (1977) in the 1970s. They observed practice choices amongst high and low rate surgeons in two regions of Scotland. They found that low rate surgeons chose watchful waiting more often and tended to base their decisions on clinical history rather than immediate physical examination, the practice of high rate surgeons. They concluded that Glover's conclusion were correct and that practice variation 'can be attributed to differences amongst specialists in their assessment practices: local difference in nature of specialist practice "create" local difference in surgical incidence'; that is, a primary cause of clinical variation was the differences in medical opinion and not the differences in morbidity.

Wennberg and his Dartmouth Medical School colleagues have researched these issues further over the last 40 years (Wennberg and Gittelsohn, 1973). Initial studies comparing processes and outcomes in New Haven and Boston showed that the latter spent more, offered more processes of care but did not have better mortality

DOI: 10.1002/hec

outcomes. He found that, in the 1980s, the risk from hysterectomy and bypass surgery was nearly twice higher in New Haven than in Boston, whereas the risks from carotid artery surgery and hip replacement operations were greater for Bostonians. It seemed that Boston surgeons had greater faith in, for instance, carotid artery surgery, whereas their New Haven equivalents preferred aspirin. On the other hand, Boston surgeons preferred medical management of coronary heart disease and the menopause, and their New Haven colleagues preferred surgical management. This work was initially published in the Lancet (Wennberg *et al*., 1987) and subsequently in the New England Journal of Medicine (Wennberg *et al*., 1989). The Glover–Bloor–Wennberg links are summarised in Wennberg (2008).

Wennberg's work is summarised in a recent book (Wennberg, 2010) and has been taken forward by colleagues such as Fisher in an extensive literature and an 'atlas' of the US Medicare variations (Fisher *et al*., 2003a, 2003b).

Fisher (2003) has boldly estimated that if conservative, safe practices were adopted, there could be potential savings of 30% of the Medicare budget. Wennberg (2010) asserts that the saving level from reduced variations may be as high as 40% of Medicare expenditure. Similar variations and the potential of savings have been identified in the UK NHS and other country's budgets (e.g. Appleby *et al*., 2011; Atlas, 2011). However, the translation of this potential into improved efficiency remains elusive.

The long history of the analysis of variations and the failure to expedite change is epitomised by the issues of *patient safety and practice registers*. Semmelweis collected data about maternal mortality in a Vienna clinic in the late 1840s and showed that rates were high, particularly in the clinic run by himself and his juniors (Stewardson and Pittet, 2011). Despite the ridicule of his peers, as there was no knowledge of germs, he introduced hand washing, and the mortality rates declined sharply. In 2000, the US Institute of Medicine published a report, which reiterated, inter alia, the need for hand hygiene and gave birth to the current increases in patient safety investments (Kohn *et al*., 2000).

Internal disputes in the Manchester Royal Infirmary in the late 18th century led Thomas Percival to publish his 'Code of Medical Ethics' (1803), which not only laid down how patients, rich and poor, should be treated but also advocated the establishment of case registers to make performance and accountability more transparent. National case registers in England remain sparse, with those for cardiothoracic surgery, bariatric surgery and hip and knee prostheses being tardy examples of good practice. The reluctance of physicians to measure and manage clinical practice with comprehensive case records and audit insulates all healthcare systems from pressures to improve transparency and accountability.

The current focus on outcome measurement also has a long history. The 1845 Lunacy Act required all facilities in the UK to report outcome data in terms of whether patients where 'dead, recovered, relieved or unrelieved'. Hospitals, which failed to report these data, faced a fine of £2 (Lunacy Act, 1845). Florence Nightingale (1863), a nursing pioneer, advocated a similar list of outcomes to demonstrate value for money.

The current use of comparative mortality data in the NHS is a product of scandals about poor outcomes, for example, paediatric cardiac surgery (Bristol Royal Infirmary Inquiry, 2001). Since 2009, there has been investment in patient-reported outcome measurement (PROMs). This applies patient self-reported quality-of-life measurements, using EQ5D (www.euroqol.org) and disease-specific measures (Department of Health, 2009) before and after treatment. As with mortality data, the initial PROMs results show significant variations in both treatment thresholds and success in improving physical and psychological well-being.

Boston surgeon Ernest Codman focused on outcome measurement and management 100 years ago. He was excluded from practising in the Massachusetts General Hospital in 1915 for advocating the measurement of 'end points' in surgery. He set up his own hospital and, with colleagues, followed-up patients systematically after surgery. He was not financially successful but helped found the American College of Surgeons and the Hospital Standardisation Program (now the Joint Commission on the Accreditation of Healthcare).

Thus, the current policy foci of reducing variations, improving safety and quality and measuring outcomes are reiterations of age-old issues that have gripped policy makers and clinicians over centuries. The primary lessons to be learnt are that there is nothing new in the policy agenda and that successive generations of

DOI: 10.1002/hec

decisions makers have tended to ignore radical reformers to maintain the status quo and the income and power of both purchasers (insurers and public agencies) and public and private providers.

The economic downturn means that healthcare budgets are under pressure internationally. The Obama reforms are likely to be inflationary as coverage increases and with its utilisation: this is exemplified, for instance, by the effects of Medicaid extension in Oregon (Baicker and Finkelstein, 2011). Given the US fiscal situation, it is imperative to control costs and improve productivity. In the UK, austerity has produced plans for stationary funding of the NHS for the next 4 years as demand continues to grow because of technological change and demographics. The government's belief is that this increased demand can be met by annual increases in productivity of over 4%.

One consequence of the pressures for higher productivity in health care is that P4P is dominating the policy agenda, even though its definition is poor and the evidence base is incomplete. The need for radical reform to improve healthcare productivity is urgent. As ever, the risk is that poorly designed, implemented and evaluated changes will worsen cost inflation and inefficiency as evidence-based policy making remains elusive. Is the health policy world clutching at straws or can P4P policies be fine tuned to increase productivity cost-effectiveness?

3. CASE STUDIES OF PAYMENT-FOR-PERFORMANCE SCHEMES AND THEIR EFFECTS

In an effort to improve transparency and performance in primary care, the UK government introduced the Quality and Outcomes Framework (QOF). This established process standards of surveillance for patients with chronic diseases. Performance was measured in terms of successful periodic review and control of conditions such as high blood pressure and diabetes. Performance was rewarded with points for successively high levels of coverage, and each point had a monetary value. The QOF bonuses were paid to each GP practice, thereby incentivising collective behaviour.

The effects of the QOF were to improve mean performance and reduce dispersion (Doran *et al*., 2006). However, there were some problems with it. It was criticised for adopting some process measures that were not evidence based in terms of improving outcomes (Fleetcroft and Cookson, 2006). The blood pressure target has been shown to have been achieved before the implementation of the QOF; that is, benchmarking was poor, and practitioners were paid for what they were already doing (Seramuga *et al*., 2011). The policy was implemented with 'light-touch regulation'. With target measurement allowing 'exemptions' when patients failed to respond to call and recall invitations to attend for care clinics. There is evidence that this light-touch regulation facilitates gaming, that is, maximising exemptions to ensure QOF payment (Gravelle *et al*., 2010).

The US Premier Hospital Quality Incentive Demonstration programme started in October 2003 and covered 267 hospitals providing care for Medicare patients. Targets were set for five conditions: acute myocardial infarction, heart failure, pneumonia, coronary artery bypass surgery and hip and knee replacements. For each condition, a series of process and outcome targets were set, for example, acute myocardial infarction has eight process targets and mortality as measures of success.

Small bonuses of 2% and 1% above normal tariff were paid to the top two deciles in terms of best performance in 2004 and 2005. From 2006, hospitals that performed below a threshold level were penalised. In the first 5 years, bonuses of $48 million were paid. A similar scheme has also been adopted in the UK NHS North West Strategic Health Authority.

Evidence of the effectiveness of the US Premier scheme is not overwhelming as there has been no randomised controlled trial and early published reports lack any control group. Studies with nonequivalent control groups have reported modestly improved quality of care scores in the participating hospitals compared with nonparticipants (Grossbart 2006; Lindenauer *et al*., 2007) and converging hospital performance, with improvement of only 1.9% in the highest performing hospitals but 16.1% in the lowest, presumably as they strove to avoid potential tariff penalties (Lindenauer *et al*., 2007). However, the improvements were based

DOI: 10.1002/hec

largely on process measures. A study found no evidence of effect on mortality or on costs (Ryan, 2009). A systematic review of all hospital P4P schemes, not just Centers for Medicare & Medicaid Services/Premier, found little formal evaluation and methodological flaws in most of the eight published studies they located (Mehrotra *et al*., 2009).

A comparison of the performance of these Premier hospitals with 780 matched controls not part of Premier showed that in the first 3 years, Premier hospitals had better results for the five conditions targeted. However, after 2007 and 2008, there was no statistical difference in the performance of the two groups (Werner *et al*., 2011).

4. PAY FOR PERFORMANCE—RESEARCH QUESTIONS

4.1. Whose performance?

Payment-for-performance incentives can be targeted at consumers, individual providers (doctors, as 'captains of the team' (Fuchs, 1974)), organisations (e.g. wards or general practices) or institutions (e.g. hospitals).

Payment-for-performance initiatives aimed at individual patients have had some success (Sindelar, 2008), and ambitious innovations are underway, for example, P4P for drug treatment (Maynard *et al*., 2011). P4P incentives to reduce obesity have also had some modest success (Cawley and Price, 2011).

However, the focus of this review is the use of P4P to improve the performance of doctors and hospitals. In this context, the pertinent question is what is the relative effectiveness and cost-effectiveness of using incentives at the level of the institution, the clinical team and the individual practitioner?

There is some evidence that the best focus of incentives is the clinical team. For instance, the NHS incentive scheme for general practitioners in primary care was conditional on practice or team performance and improved activity performance.

A counter-argument is that P4P incentive schemes should be focused on the institution rather than the individual physician as that is where the financial risk lies (Trisolini in Cromwell *et al*., 2011). This is true if institutional budgets are not devolved to clinical teams. If there is budget devolution, teams may respond positively to clinical and financial pressures to improve the efficiency of patient care.

4.2. What performance?

To measure success or failure, performance benchmarking is often used, predominantly with process measures, for example, measuring and controlling blood pressure and the use of aspirin and beta blockers after myocardial infarction. Process measures such as these are complemented by mortality data.

Ideally, the process measures used in P4P should have evidenced effects on patients' outcomes, that is, their length and quality of life. This is not always the case, for example, a critique of the NHS QOF has showed that some of the process measures used were poorly related to outcomes (Fleetcroft and Cookson, 2006), and as a consequence, the QOF's evolution is now informed by advice from the National Institute for Health and Clinical Excellence.

An obvious possible development for the QOF would be to incorporate PROMs. This might improve diagnosis if completed before consultations and would facilitate longer-term scrutiny of the course of chronic diseases.

In England, the evolution of the QOF is complemented by manipulation of tariff systems that can be used to incentivise change. The traditional hospital tariff system sets prices equal to the average cost of a procedure as in diagnostic-related group systems and what the English call 'payment by results' (PbR) (in fact, payment for activity).

The PbR system is being managed with the aim of inducing changes in practice and appears to have led to the movement of costs towards the average. Complementary initiatives have encouraged alterations in practice, for example, to accelerate the use of day case procedures for gall bladder removal, a higher tariff is paid if 70% of a hospital's procedures are treated on a day case basis.

DOI: 10.1002/hec

As in the US Medicare, the PbR system refuses payment for a group of 'never events' such as leaving foreign material in patients after operations as part of a P4P policy called Commissioning for Quality and Innovation. The PbR tariff system is also being reduced by annual uplifts below cost inflation and by a two-part tariff for emergency procedures, which reimburses hospitals at full tariff for the 2008 and 2009 volume of activity but offers only 30% of tariff for activity above that level.

In principle, the bundling of care and payment related to integration of patient pathways across primary and secondary care could be incentivised by P4P. The challenge is whether such bundling can be designed and priced efficiently.

A myriad of performance measures are being used in P4P systems. There is also a variety of payment algorithms in use including 'all or nothing' attainment targets and 'rate of improvement' target payments, which are sometimes constrained and sometimes continuous (Cromwell in Cromwell *et al*., 2011). A 'thousand flowers' are blooming in the P4P landscape.

4.3. Financial or nonfinancial incentives?

The effects of financial (bonuses and penalties) and nonfinancial incentives (reputation and peer pressure) are difficult to separate.

An important problem with the UK QOF (and the US Premier hospital programme) is whether the behaviour change was a product of the bonuses paid or the comparative performance measurement that affects the reputation of clinicians and institutions. It has been shown that hospital performance reports affect the behaviour of providers (Hibbard *et al*., 2005). The issue is whether financial and nonfinancial effects are complements or substitutes and how these effects can be separated and quantified.

The potential for poorly designed P4P incentives to erode motivation is considerable. Confucius argued that if a ruler had weapons, food and trust and was in difficulties, he should give up the first two to survive because 'without trust, we cannot stand' (Confucius, quoted in O'Neill, 2002).

Adam Smith also emphasised the role of nonpecuniary rewards, in particular, duty:

> *Those general rules of conduct when they are fixed in our mind of habitual reflection are of great use in correcting the misrepresentations of self love concerning what is fit and proper to be done in our particular situation. The regard of those rules of conduct, what is properly called a sense of duty, is a principle of greatest consequence in human life, and the only principle by which the bulk of mankind are capable of directing their actions (Smith, 1759).*

Those designing P4P incentive systems should ensure that not only can the relative effects of financial and nonfinancial interventions be identified but also that their reforms enhance and do not erode nonpecuniary incentives such as duty, trust and reputation.

4.4. Penalties and/or bonuses

Adam Smith, the original behavioural economist, noted the importance of loss aversion in human decision making:

> *Pain. . .. is in almost all cases a more pungent sensation than the opposite and correspondent pleasure. The one almost always depresses us much more below the ordinary, or what might be called the natural state of our happiness, than the other ever raises us above it (Smith 1759, quoted in Ashraf et al., 2005, pp. 176–177).*

This has been formalised and emphasised more recently (Kahneman and Tversky, 1979). Modern research has supported Smith's and Kahnemann and Tversky's contentions in a range of markets (Ashraf *et al*., 2005).

DOI: 10.1002/hec

What evidence do we have about the relative effectiveness and cost-effectiveness of penalties and bonuses in P4P schemes?

It is surprising that although penalties are part of the Premier system, there appears to be no analysis of their effects. One reason for the focus on bonuses may be that it is easier to get provider involvement in such schemes as there is no threat to income flows and the possibility of gains (Trisolini in Cromwell *et al*., 2011).

There is obvious difficulty in separating the effects of penalties from bonuses in a programme such as Premier—if a bonus is assumed, a threat of not achieving it would be perceived as a penalty. However, a crucial policy issue is whether loss aversion is the powerful incentive Smith and modern authors asserted and whether programmes using penalties are more cost effective than those using bonuses.

4.5. The size of incentives

An analysis of the Premier programme concluded that larger incentives produced a greater effect than smaller incentives (Werner *et al*., 2011). In the US experiments, to date, P4P bonuses have been less than 5% additional to standard tariff. There is advocacy of higher tariffs of 10% (Greenwald in Cromwell *et al*., 2011). Yet to be determined by evaluation is the range of values of bonuses where diminishing returns become evident.

Evidence from the Premier programme (Werner *et al*., 2011) showed that larger incentive payments produced greater improvements in performance and that better adherence to process systems and the monitoring of mortality was associated with less competition and good financial conditions.

4.6. Duration of effect

Targeting particular processes and outcomes elicits change. But is that effect permanent and after how long can bonuses be shifted to other aspects of care without any decline in the initially targeted activities? An evidence base is absent to inform change, but hopefully, analysis of pragmatic policy development will illuminate this issue.

The further development of incentives in an age of acute fiscal pressure and at best flat funding of the NHS might have a significant opportunity cost in terms of crowding out nonincentivised items of care. This may make choices as to when to shift incentives across activities even more difficult, with evidence of effect from an age of ‘plenty’ not necessarily being relevant for a period of ‘want’.

4.7. Effectiveness and cost-effectiveness

The design and implementation costs of P4P schemes are considerable. Agreeing and benchmarking the process and outcome measures together with the transaction costs for individuals and organisations and the cost of bonuses make these innovations expensive. These costs should be used to determine the relative cost-effectiveness of competing P4P programmes.

Evidence from the Premier programme indicates that competitive forces of emulation, presumably associated with retention of market share and reputation, acted as a catalyst for system change. This innovative P4P programme is relatively strong on the measurement of effectiveness but offers few cost estimates and no cost-effectiveness analysis. In addition to the costs of bonuses, there is the nice issue of the costs of management of the programme in terms of improvements in information systems and ensuring clinical teams are trained and motivated to pursue the targets. The same issues need to be measured and evaluated in relation to the effects of P4P programmes such as the QOF, Commissioning for Quality and Innovation and PbR tariff systems: what is the relative cost-effectiveness of these programmes?

A pertinent issue when exploring cost-effectiveness is that of other opportunity costs: what health and process gains, if any are given up in the nonincentivised areas of hospital activity? An evaluation of the NHS GP QOF concluded that these costs were negligible (Doran *et al*., 2011). However, this result may be explained by the generosity of the funding of the QOF in a period of rapid NHS expenditure growth. This enabled practices to hire more staff, in particular, nurses. In an age of austerity, with a hard budget constraint, displacement effects may be more significant.

DOI: 10.1002/hec

5. CONCLUSIONS: UNDERMINING INCENTIVE-INDUCED INERTIA

The curious nature of research into P4P, even when led by economists, is the focus on the measurement of whether or not it induces change—essentially, a measure of effectiveness. This replicates the myopia of medical research, which was repudiated by economists and pioneers such as Cochrane 40 years ago (Cochrane, 1972; Williams, 1972; Maynard and Chalmers, 1997). The medical myopia was epitomised by the 'evidence-based medicine' movement (Sackett and Rosenberg 1995; Maynard, 1997), which focused on encouraging practitioners to deliver those interventions that were demonstrably effective in improving patient health.

Although the measurement of effectiveness in medical care is essential, it is like a cart without a horse if it is not matched up with cost data, which demonstrates how much care is given up when a procedure is adopted. The problem now is that the literature on P4P, with its focus on effectiveness alone, does not inform policy choices in terms of the relative cost-effectiveness of competing interventions that may improve efficiency. Economists appear to have contracted a once prevalent and still common medical myopia!

Inefficiency inherent in all healthcare systems is a product of existing incentives that preserve provider incomes and give decision makers few rewards and high costs from addressing the deficiencies exhibited by providers for many decades. Inertia in reform preserves the status quo and creates expenditure pressure in public and private healthcare systems.

Payment-for-performance incentives, defined broadly, are increasingly being used to enhance competitive pressures and induce decision makers to improve their performance or lose reputation or financial rewards. It is essential that these efforts continue. Equally, it is essential that these efforts are well designed and executed so that they evidence not only the effectiveness but also the cost-effectiveness of these investments. The pursuit of this knowledge will gradually illuminate the relative efficiency of bonuses and penalties, reputational and financial incentives and other issues explored in this essay.

As further research illuminates the costs and benefits of P4P, it is important to note Mao Tse-Tung's advice:

> *Knowledge is a matter of science, and no dishonesty or conceit, whatsoever is permissible. What is definitely required is the reverse: honesty and modesty.(Mao Tse-Tung, 1966, p. 310)*

Until the unanswered questions outlined here are addressed, investors in P4P should proceed with caution, honesty and modesty, investing in robust evaluation that identifies both the costs and benefits of change.

ACKNOWLEDGEMENTS

My thanks for the comments from my editorial colleagues and, particularly, to Dr Karen Bloor for her analytical and organisational insights, which I have exploited ruthlessly!

REFERENCES

Appleby J, Raleigh V, Frosini F, Bevan G, Gao H, Lyscom T. 2011. Variations in health care: the good, the bad and the inexplicable. King's Fund. Available at http://www.kingsfund.org.uk/publications/healthcare_variation.html [accessed 2 August 2011]

Ashraf N, Camerer CF, Loewenstein G. 2005. Adam Smith, behavioural economist. *Journal of Economic Perspectives* **19**(3): 131–145.

Atlas VPM. 2011. Atlas de Variaciones en la Práctica Médica en el Sistema Nacional De Salud. http://www.atlasvpm.org/ [accessed 2nd August 2011].

Baicker K, Finkelstein A. 2011. The effects of Medicaid coverage: learning from the Oregon experiment. *New England Journal of Medicine* **365**: 683-685.

Bloor MJ, Venters GA, Sampier, ML. 1977. Geographical variations in the incidence of operations on tonsils and adenoids. *Journal of Laryngology Otol* **92**: 791–801(part 1) and **92**: 883–5 (part 2)

Bristol Royal Infirmary Inquiry. 2001. Learning from Bristol. Cm 5207, London.

DOI: 10.1002/hec

British Medical Journal Evidence Centre. 2011. *Clinical Evidence Handbook*. BMJ Publishing Group: London.
Cawley J, Price JA. 2011. Outcomes in a program that offers financial rewards foe weight loss. In *Economics of Obesity*, Grossman M, Mocan N (eds). National Bureau of Economic Research, University of Chicago Press: Chicago and London.
Cochrane AL. 1972. *Effectiveness and Efficiency: Random Reflections on Health Services Research*. Nuffield Provincial Hospitals Trust: London.
Cromwell J, Trisolini MG, Pope GC, Mitchell JB, Greenwald LM (eds). 2011. *Pay for Performance in Health Care: Methods and Approaches*. Research Triangle Press: North Carolina.
Culyer AJ, Rawlins MD. 2004. National Institute for Clinical Excellence and its value judgments. *BMJ* **329**: 224.
Department of Health. 2009. Guidance on the routine collection of Patient Reported Outcome Measures (PROMs). Available at http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_092647 [accessed 2 August 2011].
Doran TT, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M. 2006. Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine* **355**(4): 375–384.
Doran T, Kontopantelis E, Valderas JM, Campbell S, Roland M, Salisbury C, Reeves D. 2011. Effect of financial incentives on incentivised and non-incentivised clinical activities: longitudinal analysis of data from the UK Quality and Outcomes Framework. *British Medical Journal* **342**: d3590.
Ellis J, Mulligan I, Rowe J, Sackett DL. 1995. Inpatient general medicine is evidence based. *Lancet* **346**(8972): 407–410.
Fisher E. 2003. Medical care: is more always better? *New England Journal of Medicine* **349**: 1665–1667.
Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. 2003a. The implications of regional variations in Medicare spending. Part 1: the content, quality, and accessibility of care. *Annals of Internal Medicine* **18**; 138: 273–287.
Fisher ES, Wennberg DE, Stukel TA, Gottlieb DJ, Lucas FL, Pinder EL. 2003b. The implications of regional variations in Medicare spending. Part 2: health outcomes and satisfaction with care. *Annals of Internal Medicine* 138: 288–298.
Fleetcroft R, Cookson R. 2006. Do the incentive payments in the new NHS contract for primary care reflect likely population health gains? *Journal of Health Services Research & Policy* **11**: 27–31.
Fuchs VR. 1974. *Who Shall Live? Health, Economics and Social Choice*. Basic books: New York.
Glover JA. 1938. The incidence of tonsillectomy in school children. *Journal of the Royal Society of Medicine* **31**: 1219–1236.
Gravelle H, Sutton M, Ma A. 2010. Doctor behaviour under a pay for performance contract: treating, cheating and case finding? *The Economic Journal* **120**: 129–156.
Grossbart S. 2006. What's the return? Assessing the effect of "pay for performance" initiatives on the quality of care delivery. *Medical Care Research and Review* **63**: 29S–48S
Hibbard JH, Stockard J, Tusler, M. 2005. Hospital performance reports: impact on quality, market share and reputation. *Health Affairs* **24**(4): 1150–1160.
Hsiao WC. 2011. State based single payer health care: a solution for the United States? *New England Journal of Medicine* **364**(13): 1188–1190.
Kahneman D, Tversky A. 1979. Prospect theory: an analysis of decision making under risk. *Econometrica* **42**(2): 263–291
Kohn LT, Corrigan JM, Donaldson MS (eds). 2000. *To Err is Human*. Institute of Medicine: Washington, DC.
Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, Bratzler DW. 2007. Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine* **356**: 486–496.
Lunacy Act. 1845. Vic 89, House of Commons, London.
Mao T-T. 1966. *Quotations from Chairman Mao Tse-Tung*. Foreign Languages Press: Peking.
Maynard A. 1997. Evidence based medicine: an incomplete method for informing treatment choices. *Lancet* **349**: 126–128.
Maynard A, Chalmers I (eds). 1997, *Non-Random Reflections on Health Services Research*. BMJ Publishing Group: London.
Maynard A, Street A, Hunter R. 2011. Using 'payment by results' to fund the treatment of dependent drug users-proceed with care! *Addiction* **106**(10): 1725–1729.
Mehrotra A, Damberg CL, Sorbero MES, Teleki SS. 2009. Pay for performance in the hospital setting: what is the state of the evidence? *American Journal of Medical Quality* **24**: 19–28.
Nightingale F. 1863. *Some Notes on Hospitals* (3rd edn). Longman, Green, Longman, Roberts and Green: London.
O'Neill O. 2002. A question of trust. Reith Lectures. Available at http://www.bbc.co.uk/radio4/reith2002/ [accessed 2 August 2011].
Percival T. 1803. *Medical Ethics or a Code of Institutes and Precepts Adopted to the Professional Conduct of Physicians and Surgeons*. S. Russell: Manchester.
Reinhardt UE. 1982. Table manners at the health care feast. In *Financing Health Care: Competition Versus Regulation*, Yaggy D, Anylan WA (eds). Ballinger: Cambridge, Mass.
Ryan AM. 2009. Effects of the Premier Hospital Quality Incentive Demonstration on Medicare patient mortality and cost. *Health Services Research* **44**: 821–842.
Sackett DL, Rosenberg WMC. 1995. On the need for evidence based medicine. *Health Economics* **4**: 249–254.

DOI: 10.1002/hec

Seramuga B, Ross-Degnan D, Avery AJ, Elliot RA, Majumdar SR, Zhang F, Soumerai SA. 2011. Effect of pay for performance on the management and outcomes of hypertension in the United Kingdom: interrupted time series study. *British Medical Journal* **342**: d108.
Sindelar JL. 2008. Paying for performance: the power of incentives over habits. *Health Economics* **17**(4): 449–451.
Smith A. 1759. *A Theory of Moral Sentiments*. Penguin: London.
Stewardson A, Pittet, D. 2011. Ignac Semmelweis: celebrating a flawed pioneer of patient safety. *Lancet* **378**(9785): 22–23.
Syverson C. 2011. What determines productivity? *Journal of Economic Literature* **XLIX**(2): 326–365.
Wennberg JE. 2008. Commentary: a debt of gratitude to J Alison Glover. *International Journal of Epidemiology* **371**: 26–29.
Wennberg JE. 2010. *Tracking Medicine: A researcher's Quest to Understand Health Care*. Oxford University Press: New York
Wennberg J, Gittelsohn A. 1973. Small area variations in health care delivery: a population based health information system can guide planning and regulatory decision making. *Science* **182**: 1102–1108
Wennberg JE, Freeman JL, Culp WJ. 1987. Are hospital services rationed in New Haven or over-utilised in Boston? *Lancet* **I**(8543): 1185–1188.
Wennberg JE, Freeman JL, Shelton RH, Baubolz TA. 1989. Hospital use and mortality among Medicare beneficiaries in Boston and New Haven. *New England Journal of Medicine* **321**: 1168–1173.
Werner RM, Kolstad JT, Stuart EA, Polsky D. 2011. The effect of pay-for-performance in hospitals: lessons for quality improvement. *Health Affairs* **30**(4): 690–698.
Williams A. 1972. Cost benefit analysis: bastard science and/or insidious poison in the body politick. *Journal of Public Economics* **1**(2): 199–225.

DOI: 10.1002/hec