

Payment for Performance (P4P): International experience and a cautionary proposal for Estonia

**Health Financing Policy Paper,
Division of Country Health Systems**

**By: Alan Maynard
2008**

Address requests about publications of the WHO Regional Office for Europe to:

Publications

WHO Regional Office for Europe

Scherfigsvej 8

DK-2100 Copenhagen Ø, Denmark

Alternatively, complete an online request form for documentation, health information, or for permission to quote or translate, on the Regional Office web site (<http://www.euro.who.int/pubrequest>).

© **World Health Organization 2008**

All rights reserved. The Regional Office for Europe of the World Health Organization welcomes requests for permission to reproduce or translate its publications, in part or in full.

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by the World Health Organization in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by the World Health Organization to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either express or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall the World Health Organization be liable for damages arising from its use. The views expressed by authors, editors, or expert groups do not necessarily represent the decisions or the stated policy of the World Health Organization.

About the author

Alan Maynard is a Professor of Health Economics at the University of York, England (e-mail: akm3@york.ac.uk), and currently Visiting Adjunct Professor in the Centre for Health Economics Research and Evaluation (CHERE) at the University of Technology in Sydney, Australia. In addition to his academic work, he has been involved in NHS management for 25 years and since 1997 has been chairman of the NHS Hospitals Foundation Trust Hospital which has a budget of £170mn and 650 acute beds, serving a population of 350 000. He is a member of the English Department of Health's External Advisory Board for Payment by Results (DRGs) and is currently a Specialist Adviser to the Select Committee on Health of the House of Commons (United Kingdom Parliament).

Acknowledgements

The author would like to thank Tamas Evetovits and Jarno Habicht from WHO, and Mari Mathiesen, Helvi Tarien, Triin Habicht, Kristiina Kahur and Jane Alop from EHIF for their comments on the earlier versions of the paper. Discussions during the workshop "Payment Methods in Health Care - International Developments and Challenges for Estonia" held on 4 December 2007 in Tallinn were also helpful.

CONTENTS

	<i>Page</i>
Executive summary	1
Background.....	3
Common Problems confronting purchasers and providers.....	4
Introduction: focus on structure, process and outcome	4
Common problems in the delivery of health care.....	5
Incentivizing change in health care provision and purchasing	14
The importance of incentives	15
Paying hospitals	15
Paying doctors	17
Reinforcing the purchaser role: the development of normative incentives	18
Normative incentives for hospitals	19
Normative incentives for doctors	22
Conclusions.....	24
References.....	27

Executive summary

1. Estonia has established an effective organizational structure to finance and provide health care for its population. The system has advanced contracting and provider payment systems and additional incentives to enhance quality are desired.
2. However, in common with the rest of Europe and other middle and high income countries, well researched deficiencies in health care delivery have been encountered, in particular:
 - a. an incomplete evidence base about ‘what works’ in medicine, with the majority of interventions lacking an evidence base;
 - b. large variations in clinical practice with patients with similar characteristics getting very different care;
 - c. the delivery of inappropriate care with no benefit to patients;
 - d. failure to deliver appropriate, evidence-based care to patient groups such as the chronically ill;
 - e. hospital error rates of perhaps 10% of all admissions, based on projections from international data; and
 - f. an absence of patient-reported outcome measurement.
3. These deficiencies have been evident for decades but reforms of structure and process have had little impact on them. Worldwide, and particularly in the United States and England, this has led to the gradual deployment of process performance indicators, increasingly with incentives at the margin, small financial gain and losses for hospitals and specialists: paying for performance (P4P).
4. Examples of P4P are:
 - a. CMS – Premier Inc. in the United States where a battery of measures have been adopted for five clinical conditions (acute myocardial infarction, knee and hip replacement, pneumonia, heart failure and coronary artery bypass grafts) in the Medicare system (largely elderly patients). Performance in relation to these indicators creates annual rewards of 1% and 2% of Medicare DRG revenue, and after three years, penalties. Uncontrolled evaluation indicates a significant improvement in the processes of care in these areas.
 - b. CMS has also imposed DRG revenue reductions if contracted hospitals fail to ‘volunteer’ to collect a uniform set of performance data that can be used to monitor performance. The use of such common data sets is now routine in England.
 - c. Medicare has announced that from October 2008 it will not reimburse the cost of treating medical errors such as pressure sores and catheter induced infections.
 - d. In England, fines are being introduced for hospitals who fail to meet reduction targets for Methicillin-resistant *Staphylococcus aureus* MRSA and *Clostridium difficile* infections.
 - e. These institutional initiatives are being complemented by reforms adopting hybrid systems of payment to medical practitioners, e.g. small losses and gains in income in relation to process and activity performance.

- f. In England, and in some hospitals in the United States, mortality outcome data is being supplemented with patient-reported outcome measures (PROMs) using standardized instruments (in thousands of clinical trials) to evaluate any gains in physical and psychological functioning. Complementary P4P process measures with PROMs are experimental but an essential step in determining whether spending millions of euros on health care actually makes patients better.
- 5. P4P reforms have to be carefully designed, implemented and evaluated. Piloting is essential to determine local feasibility and ‘fine tune’ instrumentation and measurement.
- 6. Incentives such as P4P can be very powerful in their effects. Caution in their design and deployment is essential. However, ignoring their potential would be unwise, as they offer the possibility of improving value for money for taxpayers and patients.

Background

Estonia is a high-income country that has developed a robust system of public financing of health care. The Estonian health system has developed solid mechanisms for collecting funds, and the Estonian Health Insurance Fund (EHIF) is well-established and offers an efficient mechanism for pooling resources. A key function of EHIF is purchasing of health care services for insured population. Its role as a purchaser of health care makes it relevant to this paper, which was commissioned by World Health Organization and EHIF, as part of the Biennial Collaborative Agreement between WHO and the government of Estonia. The terms of reference describe the aim of this work as “to produce a conceptual paper describing the scope, potential use, alternatives to and limitations of pay for performance (P4P) for providers, particularly concentrating on the hospital sector, and analysing its adaptability to the health care system in Estonia”. The paper will describe and analyse international experience and its relevance to Estonia, but it is not directly applicable to the Estonian health system. The overall objective is to facilitate further country-specific policy discussions in Estonia building on the international evidence.

The hospital sector is the primary focus of this report. However, performance management of the hospital sector has two overlapping elements. First, managers at the hospital level can be given financial and non-financial incentives to improve performance in relation to adoption of new technology, reducing length of stay and reducing performance variability. This requires managers to negotiate change and to achieve compliance with specialist clinicians, whose influence on resource allocation often dominates behaviour and performance. Second, P4P incentives can be designed and targeted directly at clinical specialists who are then obliged to work with management to comply with EHIF purchaser mandates. A hospital may hit its performance targets but clinical practice within the facility may exhibit unacceptably large variations in clinical behaviour that merit targeted contracting to mitigate internal inefficiency. For instance, a hospital might hit its performance targets but the use of day surgery by its clinicians may vary unacceptably in relation to the evidence base. Behaviour such as this might merit P4P incentives targeted at specialists in addition to those targeted at the hospital.

The first section of the paper will review evidence of common provider problems in all health care systems and their implications for introducing a P4P system. This will be followed by a review of P4P reforms in the United Kingdom and United States in particular. Throughout this analysis, there is an emphasis on the incompleteness of the evidence base and the need for careful experimentation and evaluation to inform service reform.

The ultimate goal of P4P reforms is to promote evidence-based care and to ensure that in doing so process management is supplemented by patient-reported outcome measures (PROMs). It is essential to gradually and systematically shift the policy processes from the mere analysis of process or “doing things to patients” to outcome measurement and management to inform decision-makers about whether health care expenditure actually “make patients better”.

Common Problems confronting purchasers and providers

Introduction: focus on structure, process and outcome

Over 40 years ago Donabedian emphasized the distinction between structure, process and outcome (Donabedian, 1966). The focus of reformers is usually on the first two items in Donabedian's list. Public and private agencies confronted by policy difficulties tend to "move the deck chairs on the Titanic" by "redisorganizing" the health care delivery structures. Implicit in this response is the usually evidence-free belief that changes in structure will somehow improve in some implicit way both processes and outcomes. This optimism is usually unjustified and brings to mind the alleged remarks of an administrator for the emperor Nero, to the effect that reorganization was a wonderful method of giving the illusion of progress, while in fact producing confusion, inefficiency and demoralization.

Structural reform is often complemented by initiatives to improve processes. Processes are means to an end, namely improvement of patient-reported quality of life. However, process reform usually takes place without consideration of outcome. It is generally assumed that providing more primary care consultations, more hospital outpatient and inpatient care and less waiting time are beneficial for patients. This outcome should be proven with measurement rather than assumed. Instead, health care policy-makers continue to reform structure and processes, and devise sophisticated process performance measures, while failing to assure themselves that their innovations are successful in improving patient health.

In some (perhaps too few) cases, improvements in processes may improve outcomes. Excellent examples of this are efforts to reduce medical errors and improve patient care and outcomes with improved hygiene. However appealing such investments may be, they should be informed by cost-effectiveness information. Typically such investment initiatives are vaguely conceived with incomplete evidence to support design and implementation, and little evaluation of success in terms of improved processes and patient outcomes.

A focus on process also creates simple questioning of health care policies with high public profiles such as waiting time. How much health gain has been produced by reductions in waiting time in the English National Health Service (NHS)? In many countries waiting time for elective procedures is politically highly sensitive, but perhaps those who wait are adjudged by specialists to have little scope for health gain and consequently are not prioritized. Such issues are empirical matters: where is the evidence to support the contentions of the competing sides of the waiting time argument? Without such evidence, politicians find it difficult to manage public indignation over a phenomenon that may be inefficient to drive down to zero.

Such debates are relevant in eastern Europe as well as the United Kingdom. For instance, even though the clinical evidence of effectiveness may be incomplete, political expectations in Estonia and England may dictate investment in cancer and other areas of care, particularly when there is evidence that survival rates are inferior those of France, Germany and the United States. While political prioritization is inevitable, it should be informed by evidence when possible. When there is no evidence available, investment in evaluation should be a priority, although it is often low on the political agenda. After all, the objective is to target funding at those patients who can get the greatest health care gain at least cost.

In such a debate the crucial issue is measurement of patient outcome, the third and relatively neglected category in Donabedian's list. Asking if the patient is better after care raises some

interesting issues. First, whose view should count? There is good evidence that expert opinion is an incomplete way of answering this question. The central person in making this judgement should be the patient. However, in looking to the patient to determine whether health care improves patient health, efforts have to be made to ensure that the focus is on outcome rather than process. There is a risk that patients' focus will be on waiting times and on other investments that are poorly evidence-based and subject to marketing by myopic provider groups (e.g. the pharmaceutical industry).

Furthermore, what is the meaning of "better"? Patients hope for improvement in the length and quality of life. Measurement of the quality of life requires patient assessment of physical and psychological functioning before and after a medical or surgical procedure. This is routine in clinical trials. For instance, cancer trialists have been persuaded over recent decades from merely measuring the duration of patient survival to also assessing their quality of life. To promote their (often marginally beneficial) new products, pharmaceutical companies routinely measure the quality of life of patients in order to estimate health gain in terms of quality adjusted life years (QALYs) for regulatory agencies such as the English National Institute of Health and Clinical Excellence (NICE). It is curious that what is routine practice in clinical trials is absent in routine clinical care. The measurement of outcomes has been under-developed despite specific and generic measures of the quality of life having been available for decades and used routinely in clinical trials (Stewart, Ware, 1992; Patrick, Erickson, 1992). Yet without investment in such measures in routine health care, how can policy makers determine whether patients get better and health care investments improve patient well-being?

Common problems in the delivery of health care

Regardless of the public-private mix in the local health care system there are some common and well researched problems that have been largely ignored for decades by policy makers reforming the purchasing and provision of health care (Maynard, 2005). There are five such problems and each will be discussed, with the P4P implications highlighted:

1. incomplete evidence of what is clinically effective in medicine;
2. variations in clinical practice;
3. appropriateness of service delivery;
4. patient safety; and
5. reluctance of policy-makers and providers to measure and manage "success" in medicine using patient reported outcome measures (PROMs).

It is assumed that these internationally widespread problems are also endemic to Estonia.

What works in medicine?

Archie Cochrane expressed great concern about the evidence base of medicine and the efficiency of investing in health care over 30 years ago, summarizing the problem of the relationship between inputs and outputs in an amusing comparison.

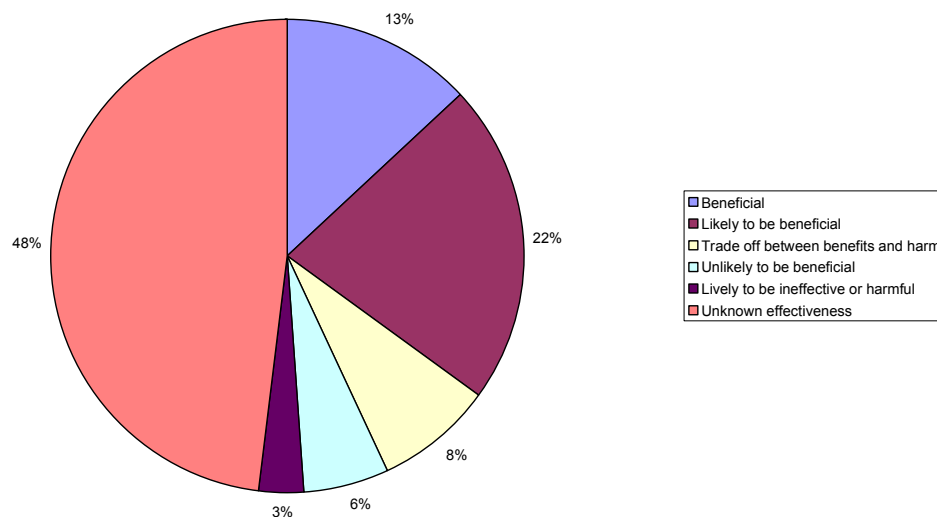
I once asked a worker at a crematorium, who had a curiously contented look on his face, what he found so satisfying about his work. He replied that what fascinated him was the way in which so much went in and so little came out. I thought of advising him to get a job in the NHS, it might increase his job satisfaction, but decided against it. He probably gets his kicks from the visual

demonstration of the gap between input and output. A statistical demonstration might not have worked so well. (Cochrane, 1972)

This is not to argue that all medicine is useless. Much has been proven to be clinically effective (Nolte, McKee, 2004). However, despite improvements in the health care knowledge base, the degree of uncertainty about the clinical effectiveness of a significant proportion of medical practice remains high.

Figure 1 shows the percentages of common medical practices that are proven and unproven. This type of analysis uses Cochrane collaboration criteria that focus on “gold standard” evidence, i.e. randomized controlled trials (RCTs). The authors conclude that over 45% of activity lacks an evidence base. This is not to say it is not clinically effective, but contends that a large amount of health care is unproven and that the uncertainty surrounding its effectiveness needs to be explored in well-designed and reported RCTs. With approximately 13% that is beneficial and 22% likely to be beneficial, over half of medical practice is unproven.

Figure 1. Uncertainty of clinical effectiveness



Source: *BMJ*, 2007

Any investment in health care needs to be informed not merely by evidence of clinical effectiveness, but also by evidence of cost-effectiveness. What is clinically effective may not be cost-effective but what is cost effective is always clinically effective. The usefulness of cost-effectiveness over mere clinical effectiveness has been accepted for many years (Maynard, 1997) and is a fundamental element of current health technology assessment, epitomized by NICE, as well as the Scottish Medicine Consortium and the Australian Pharmaceutical Benefits Scheme, to name three.

The implication of this uncertainty about what works cost effectively in medicine is that when designing P4P systems, it is essential to include incentives that reward provision of services of proven efficiency. Furthermore, the design of P4P systems has to be flexible so that if evidence emerges of lack of relative effectiveness and cost-effectiveness, technologies can be abandoned and replaced by more efficient interventions. Typically in fee-for-service payment systems it is

difficult to remove inefficient procedures from fee schedules. This is epitomized by the difficulties encountered by German policy-makers when they removed spa treatments from the payment schedule.

This problem should not be marginalized. Health technology systems tend to add new procedures to the medical armoury. However, investments in the measurement of cost-effectiveness often fail to investigate the scope for abandonment of existing technologies that have little benefit to patients (Select Committee on Health, 2008). With over 50% of everyday treatments of uncertain effectiveness let alone cost-effectiveness, it is curious that all health care systems invest so little in eroding the unproven and challenging what is claimed to be proven. Such challenges require robust governance, which may be best managed by collaborating clinicians. Investments in P4P will be low-yielding without strong regulatory input that uses the existing evidence base and funds its on-going improvement.

All markets have to be regulated; even stock exchanges must have rules that determine how ownership of assets is transferred and obligations to pay are to be enforced. Furthermore, as emphasized by such founding fathers of economics as Adam Smith, markets are always threatened by capitalists as they seek to create monopolies and other methods of increasing their profits. Public health care markets are no exceptions to such pressures and neglecting to regulate them effectively may create increased perverse effects as P4P is developed.

Variations in clinical practice

For decades researchers have reported large, unexplained variations in the delivery of care to patients with similar medical and social characteristics. Such variation is unsurprising given the uncertainty about the clinical and cost-effectiveness of much of medicine. This uncertainty permits variation in practice style and the independence of practitioners protects doctors from scrutiny and management. More remarkably, and sadly, the medical professions themselves have failed to police and manage practice variations among their members even though they usually fiercely defend their right to self-management.

The problem was described by the American physician Jack Wennberg when he argued over 30 years ago that “the amount and cost of hospital treatment in a community have more to do with the number of physicians there, their medical specialties and the procedures they prefer than the health of the population” (Wennberg, Gittelsohn, 1973). Wennberg and his Dartmouth College colleagues published studies in the 1980s highlighting variations in Medicare practice in the United States (Wennberg, Freeman, Culp, 1987; Wennberg et al., 1989). This work was extended to the measurement of variations in Europe (McPherson, 1982), and has been updated by the Dartmouth group (Fisher et al., 2002), which described outliers in terms of per capita spending on Medicare (e.g. \$10 500 in Manhattan and \$4823 in Portland, Oregon) and showed that these differences were due to volume effects and not differences in illness, socioeconomic status or the price of services. Fisher went on to question whether more spending on medical care produced more health and noted that if all practitioners could be persuaded to adopt the safe practices of conservative treatment regions, there were potential savings of 30% of the Medicare budget (Fisher, 2003).

Variations in clinical practice have been researched and highlighted by policy-makers in many other countries. For instance, the United Kingdom Department of Health and Social Security (1976), faced with zero growth in funding in 1976, described studies of best practice and noted variations. As in the United States, the British have regularly noted practice variation and the

inefficiency it produces, but have failed to manage it systematically. The current government is again pointing out tardiness in adopting more efficient technologies such as day case surgery and the large variations in clinical practice (NHS Institute for Innovation and Improvement, 2006).

An example of such variation is the activity of English hospital consultants (specialists). Administrative data on hospital activity has been collected in England since 1989, but has not been used to manage clinical practice. For every hospital admission, the Hospital Episodes Statistics (HES) record the referring general practitioner, patient characteristics, hospital procedures and outcomes (e.g., complication rates and mortality). Although collected with increasing diligence and accuracy, the data have not been used to inform management until recently. Simple analysis of such data enables managers to interrogate clinical activity more thoroughly. Figures 2 and 3 show the national distribution of general surgery consultant activity. The first figure measures volume in terms of finished consultant episodes (FCEs). The second takes these volumes and adjusts them crudely for case mix differences. The six vertical lines rank the consultants of one hospital in relation to the national distribution. It seems that some surgeons in this hospital were working hard while others may have been indulging in “on the job leisure”!

Figure 2: Variation in activity in general surgery: FCEs

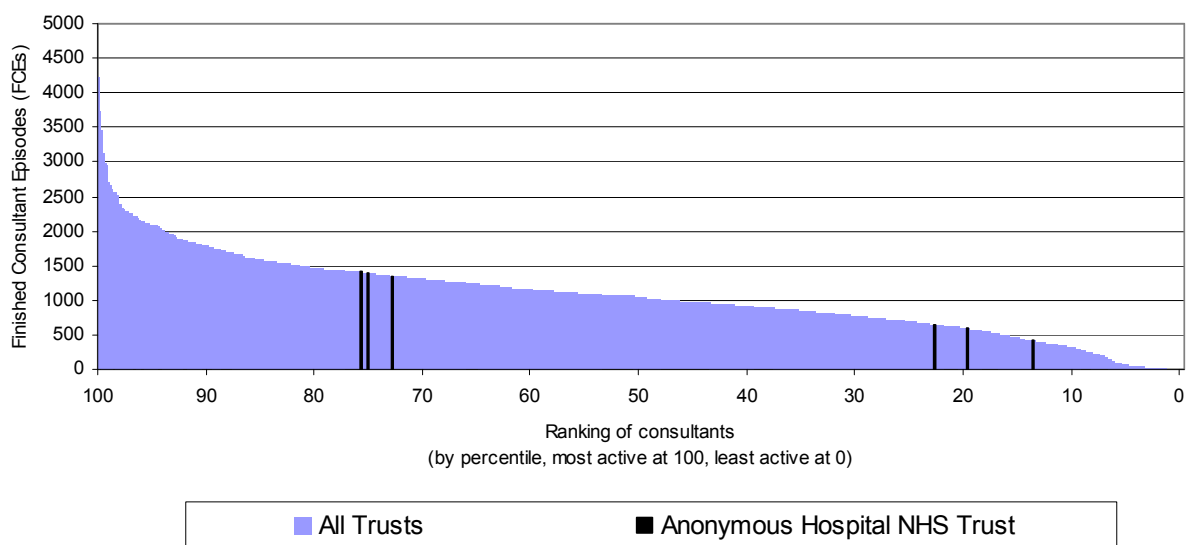
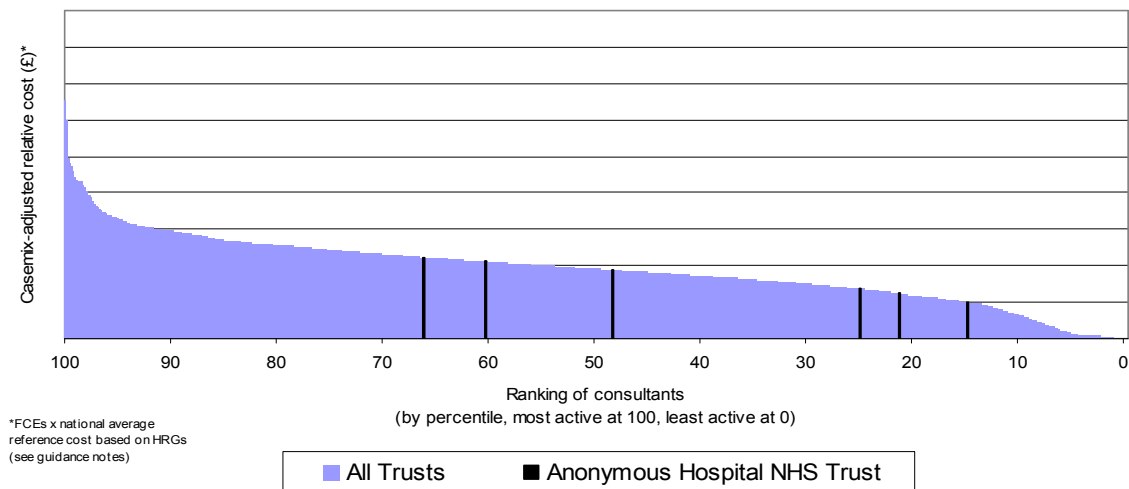


Figure 3: Variation in activity in general surgery: case mix adjusted



This simple analysis of consultant activity variations was first published and distributed by government to all NHS acute hospitals in 2002 (Bloor & Maynard, 2003a). It was repeated in 2006 (Bloor & Maynard, 2006) and in 2007 the English Department of Health finally regularized the production and distribution of these data, i.e. each English hospital now receives relative performance data for specialists in 10 medical and surgical fields. This enables management to identify outliers and work collaboratively with physicians to reduce dispersion around the mean. Simple descriptive data of this nature do not explain the activity levels of surgeons in this case: the variation may be as a result of many factors, and the relative position of practitioners tells managers nothing about patient outcome. However, it does focus managerial attention on the variations and the need to investigate them using Six Sigma® production engineering techniques, for instance. This approach focuses managerial attention on items three standard deviations above and below the mean, so that lessons can be learnt from the very good and the very poor performers.

Analysis of clinical practice variations also has implications for workforce planning. If means can be shifted and dispersion reduced, more work can be produced by the existing labour force. The tradition in health care has been not to manage process productivity systematically. If it was managed better the demand for physicians could be moderated. Fisher, in the context of the current American debate about an alleged physician workforce shortage, has asserted that the removal of 30% of American doctors to Africa could benefit Africans with no disadvantage to American patients if clinical practice variations in the country were removed (Brownlee, 2007). Such provocative assertions may help concentrate the minds of clinical and non-clinical management on the productivity potential in reducing practice variations. It also shows how a focus on process may elicit efficiency gains and the need to complement practice variation reduction efforts with careful scrutiny of patient outcomes.

The implications of variations in clinical practice for P4P are that evidence has to be provided to determine whether such incentives can be used to shift the mean rates of productivity and reduce the dispersion around the mean. Obviously such changes must be demonstrated to be cost-effective. The measures used here are of process and activity rather than PROMs. It may be that practitioners with relative low activity rates have superior PROMs compared to those with high activity rates. Of course the opposite may also be true, and there is evidence of this in some surgical specialties (Halm, Lee, Chassen, 2002, for example). Thus, using P4P incentives with a

focus on activity has to be undertaken with care, and requires complementary investment in outcome measurement.

Appropriate clinical practice and related P4P experience

In most health care systems there is evidence that proven, cost effective interventions are not delivered to patients and, more contentiously, that inappropriate interventions are given to patients with no benefit and possible harm. International clinical practice fails to deliver proven, cheap and cost-effective interventions to chronic patients, who are a major source of expenditure and their incomplete treatment produces avoidable morbidity and mortality.

A study by the RAND Corporation in the United States concluded that on average American patients received only 55% of the care needed (Kerr et al., 2004). Performance management of private and public agencies delivering chronic care in the United States seems limited. For instance the Healthcare Effectiveness Data and Information Set (HEDIS) of the National Committee for Quality Assurance (NCQA) has 71 components in 8 quality domains. The process domains are asthma control, provision of beta blockers after a heart attack, control of blood pressure, comprehensive diabetes care, breast cancer screening, antidepressant medication, child and adolescent immunization and advice to quit smoking. These familiar and well-documented areas are similar to those in many other countries. HEDIS seeks survey returns from private insurers, Medicare and Medicaid and attempts, to the extent that contributors permit publicity, to provide comparative analysis to facilitate employer and consumer choices of insurance plans. However, the data are not comprehensive and apart from publicity the incentives to improve provision are limited. As can be seen from Table 1, the measures used in HEDIS are basic indicators for relatively well-documented interventions in chronic care and prevention. However, given the poor record of American purchasers and providers in delivering these services to patients, they are a sensible way of seeking improvement.

Table 1: The Healthcare Effectiveness Data Information Set (HEDIS) 2008 measures

Effectiveness of care	
1	appropriate testing for children with pharyngitis
2	appropriate treatment for children with upper respiratory infection
3	avoidance of antibiotic treatment in adults with acute bronchitis
4	use of spirometry testing on assessment and diagnosis of COPD
5	pharmacotherapy of COPD exacerbation
6	use of appropriate medication for people with asthma
7	cholesterol management for patients with cardiac conditions
1	childhood immunization status
2	adolescent immunization status
3	lead screening in children
4	breast cancer screening
5	cervical cancer screening
6	colorectal cancer screening
7	Chlamydia screening in women
8	glaucoma screening in older adults

Source: http://web.ncqa.org/Portals/O/HESIS2008/2008_Measures.pdf

In the United Kingdom, concern about the failure to deliver chronic care to patients led to the Quality and Outcomes Framework (QOF) to motivate GPs. Initially it had 10 clinical components, with points allocated for achievement in them. Currently 1000 points can be earned,

each worth £125 for successful performance by group practices (NHS Institute for Innovation and Improvement, 2006). The content of the QOF is similar to HEDIS, but is more ambitious as can be seen in Table 2.

Table 2: United Kingdom Quality Outcomes Framework (QOF) for primary care, 2004

Disease	Performance indicator
Asthma	% of patients with asthma who have had an asthma review in the previous 15 months
Cancer	% of patients with cancer reviewed within 6 months of confirmed diagnosis
Chronic obstructive pulmonary disease (COPD)	% of patients with COPD with diagnosis confirmed by spirometry and reversibility testing
Coronary heart disease (CHD)	% of patients with CHD whose last blood pressure measurement was 150/90 mm Hg or less
Diabetes	% of patients with diabetes whose last blood pressure measurement was 145/85 mm Hg or less
Hypertension	% of patients with hypertension with last blood pressure measurement was 150/90 mm Hg or less
Hypothyroidism	% of patients with hypothyroidism with thyroid function tests recorded in the previous 15 months
Mental health	% of patients with severe long-term mental health problems reviewed in the preceding 15 months

This has been recognized as an interesting innovation by some Americans (Epstein, 2006; Doran et al., 2006) but its opportunity costs have been high in terms of payment levels and lack of evidence of which interventions were given up by practitioners in their pursuit of the QOF fee-for-service rewards. Furthermore, the intervention standards are modest. They are not well-based on evidence in some cases and the point weighting of the categories is related more to clinical workload than to potential health gain (Cookson & Fleetcroft, 2006).

In addition to these problems, there are also other general issues associated with incentives related to chronic care in P4P systems. Typically, such fee-for-service payments may be difficult to alter over time as deletion and addition of elements of care redistribute practitioners' incomes and create conservative incentives. Furthermore such systems typically induce gaming or cheating as providers legitimately and illegitimately seek to maximize their rewards. Evidence of this has begun to emerge from the recent QOF scheme. (Centre for Health Economics, 2007). As with all P4P investments, the policy challenge is how to minimize the risk of gaming with carefully targeted regulatory investments and demonstrate evidence that the benefits of offering provider incentives exceed the costs lost to cheating.

A considerable amount of work on the delivery of interventions to patients who get no benefit from them has been initiated by Professor Robert Brooks and colleagues in California. Their general approach is to use the evidence base and supplement this with expert consensus to identify appropriate practice. Having identified what should be delivered to whom, retrospective analysis of clinical choices in relation to best practice is undertaken. The use of evidence and consensus agreements by experts when applied retrospectively to actual performance reveals the delivery of inappropriate care. An example of this involved British and American panels of experts, separately devising criteria of appropriateness for coronary artery bypass grafts (CABGs) and angioplasty and then applying these criteria to the records of patients treated in the

Trent (Sheffield) region of the English NHS. The panels' evidence showed a large volume of inappropriate care.

Inappropriate care, even in the face of waiting lists, is a significant problem in Trent. In particular, by the standards of the United Kingdom panel, one half of the coronary angiographs were performed for equivocal or inappropriate reasons, and two fifths of the CABGs were performed for similar reasons. Even by the more liberal United States criteria, the ratings were 29% equivocal or inappropriate for CA and 33% equivocal or inappropriate for CABGs. (Bernstein et al., 1993)

The combined effects of failing to deliver appropriate and cost effective care, particularly for patients with chronic disease, and delivering care of little or no value to patients is again indicative of service inefficiency that should be challenged by both clinical and non-clinical managers, and improved by P4P incentives.

Patient safety

Another policy area in which the lessons of the past have been forgotten with peril to patient welfare is patient safety. The Hungarian physician Semmelweis was called the “saviour of women” for his discovery in the 1840s that puerperal fever could be avoided – and maternal mortality thus be significantly reduced – by improved hand hygiene. Since then, the need to have high performance standards for hand washing appears to have been lost, perhaps in part due to the availability of antibiotics. However, the importance of basic patient safety procedures has been rediscovered in recent years, led by Swiss research (Pittet, 2001).

The United States Institute of Medicine report *To err is human: Building a safer health care system* was shocking in its revelation that medical errors killed more citizens than breast cancer, HIV-AIDS or motor vehicle accidents (Kohn et al., 1999). The authors estimated from limited prevalence data that 44 000–98 000 Americans died each year as a result of errors in American hospitals. Limited British data has led the government to argue that the error rate in NHS hospitals is around 10%, that is, one in ten inpatients are victims of errors (Vincent, Ennis, Audley, 2004; Sari et al., 2006). These estimates are derived from limited retrospective examination of patient records. Routine reporting of errors to the National Patient Safety Agency (NPSA) typically reveals significant under-reporting compared to case record review.

The consequent issue is how to reduce errors cost effectively. A major industry has grown up to reduce medical errors, for example, the American Institute for Health Improvement (IHI)'s 5 Million Lives Campaign. Such organizations usually assert that their interventions are proven but the evidence base for clinical and cost-effectiveness remains incomplete. The challenge for P4P policy-makers is to target incentives. Just what behaviours should be targeted to reduce medical errors and improve patient safety cost effectively? For example, IHI has had a major impact on American policy and focused continuing attention on the reduction of avoidable medical errors, and European organizations such as the Health Foundation in the United Kingdom have invested in disseminating IHI practices to NHS hospitals. Yet, the interventions used are of uncertain clinical and cost-effectiveness and when IHI advocates are challenged with demands for evidence, they sometimes respond defensively, appearing to believe that their work is a good thing. But evidence is needed to determine whether the efficient level of medical errors is zero, or if there is a level of errors after which reduction is not cost-effective, for example. With the majority of errors likely to affect the elderly, who dominate hospital bed occupancy, and evidence that many victims of errors would have been dead within three months anyway, is investment in IHI practices efficient? Robust questions such as these are as essential as they are often absent when “faith” based policy advocacy dominates health care policy making. There is

now a slowly emerging consensus that investment in improved patient safety requires a much more robust evidence base (Maynard, 2007; Grol, Berwick, Wensing, 2008).

Measuring and managing patient outcomes

The 1844 United Kingdom Lunacy Act required public psychiatric hospitals to record patient outcomes as “dead, relieved or unrelieved” upon discharge. They did so throughout succeeding decades and some acute units measured success in this way until the creation of the NHS in 1948. More than 100 years ago Florence Nightingale also described patients who left her care according to this simple classification of outcomes (Nightingale, 1863). Remarkably, this early advocacy of outcome measurement and management is absent in current public and private health care systems.

President Reagan’s administration began the publication of hospital mortality rates over 20 years ago and similar disclosure policies are now quite common. However, hospital mortality rates are typically relatively low. The medical profession’s adherence to mortality as the criterion of failure is not matched by equal attention to making patients better as a standard of success. This is paradoxical, as clinical trials typically use generic quality of life measures such as Short Form 36 (www.sf36.org) and EQ5D (www.euroqol.org) as well as disease-specific measures such as VF 14, used to assess visual acuity before and after cataract removal. The two generic measures have been translated into dozens of languages and used in thousands of clinical trials. There are both Estonian and Russian language versions of them. The use of EQ5D before and after hip and knee replacements has the potential to identify improvements in physical and psychological well-being of patients. This is an essential element of any policy of consumer protection and also facilitates the identification of good providers from whom lessons can be learnt and poor providers whose practice needs to be improved.

The corollary of managing performance is measuring patient-reported outcomes. This involves not only mortality data but also the quality of patients’ well-being before and after health care. The instruments to measure patient-reported outcomes are available and well validated. If the goal of policy is to improve patients’ health, such measurement is essential and P4P policies that ignore it are incomplete. In Britain, this was first recognized by the British United Provident Association (BUPA) a non-profit private insurer. Confronted by the poor performance of a gynaecological surgeon who rendered scores of women incontinent following hysterectomy surgery, BUPA decided to measure patient outcomes as a means of improving consumer protection and of assessing the relative success of the surgeons they employed (Vallance-Owen & Cubbin, 2002). From 1999–2007 they deployed SF36 for all their patients before and three months after procedures such as hip and knee replacements, hernia repairs, hysterectomies and cataract removals. The results of this work were fed back to clinicians, with particular attention paid to performers in the tails of the SF36 distributions for physical and psychological functioning. They found that SF36 was insensitive for cataracts, and replaced it with a specific measure that assesses visual acuity (VF14), and found that a significant proportion of cataract operations yielded no improvement in visual acuity.

The implication for Estonia and other countries with data-rich systems is that they may have considerable stocks of process data but, management’s use of such material is inevitably restricted if there is no complementary patient outcome data. Reducing variations in clinical processes, for instance, brings with it the risk that some patient outcomes will worsen. To guard against such adverse effects, process reform has to be cautious and complemented with PROMs.

In 2007 BUPA sold its hospitals, and is now developing contracting so that, as the largest private insurer in the United Kingdom, it will require all providers to report PROM data as part of the accountability and reimbursement framework. This is being deployed progressively for cataracts (VF14) and hips and knees (SF12/SF36). The English NHS is now beginning a PROMs programme (see below). In the public health care system in England, a commercial consultancy, CHKS, has collaborated with four NHS acute trusts and, since June 2006, used EQ5D and SF36 for all elective patients. Early results from this work show that these instruments are useful in identifying gainers and losers. However, the response rate is modest (about half that of BUPA's 70% before and after responses).

Work by researchers at the London School of Hygiene and Tropical Medicine for the English Department of Health has also demonstrated that PROM data can be collected at a cost of around £10 per patient. This work is creating a knowledge base which is attracting policy-makers' interest. NHS reforms over the last decade have generated large rents for the labour force and modest changes in process activity (Wanless et al., 2007). The Atkinson report for the United Kingdom Office of National Statistics (Atkinson, 2005) was commissioned by a government increasingly frustrated by its inability to see a return on increasing investment in the NHS. In the 2002–2007 period expenditure on the NHS increased by 50% or some £43 billion. Government indicators of process productivity show modest if not negative changes. Consequently, the Atkinson report advises investing in NHS outcome measurement and there is an interest in government to do so. Furthermore, English regulatory agencies such as Monitor (www.monitor-nhsft.gov.uk), which supervises Foundation Trust hospitals, and the Healthcare Commission (www.healthcarecommission.org.uk), which grades the performance of all providers and purchasers in the English NHS, are both now supporters of investment in PROM.

Summary

The delivery of healthcare internationally is inefficient. Public and private sector insurers have failed to protect taxpayers and insurance contributors from the financial consequence of well-researched inefficiency. There appears to be increasing recognition of these failures and the need to contract with providers in ways that reward improved resource allocation. The existence of these problems, common to all health care systems, have clear implications for the introduction of a P4P system in Estonia. P4P has the potential to further reduce the provision of inappropriate and inefficient procedures. It is essential to measure and monitor the cost–effectiveness of P4P intervention in order to reduce variations in practice and improve outcome measurement.

The opponents of P4P may focus on the risks of perverse incentives. However, the inefficient status quo is untenable in all health care systems and failure to invest in P4P complemented by experimentation with PROMs will perpetuate the existing inefficiencies in provider arrangements in public and private health care systems. Such an outcome may, of course, favour P4P opponents in terms of their income and employment, but it would not favour patients and taxpayers. Cautious and carefully designed, implemented and evaluated change involving the application of P4P incentives is essential to reducing inefficient health care delivery.

Incentivizing change in health care provision and purchasing

Whether we take a Donabedian view of health policy or the more medical view of Cochrane, it is clear that health care reforms have focused on structure and process, and have neglected outcomes. The evidence-based medicine movement initiated by Cochrane has improved policy

and practitioner focus on what works from a clinical effectiveness perspective. This has developed nicely in medicine and led to the development of the focus of technology appraisal on both clinical and cost-effectiveness, with the latter accepted as the dominant determinant of resource allocation. However, in health policy the focus on input-output/outcome relationships remains poor, with continued undue emphasis on structure and process rather than PROM, despite research evidence, at the cost of continued clinical practice variations and failure to apply the quality of life measures used extensively in clinical trials to routine medical practice.

The importance of incentives

Incentives can induce behaviours consistent with policy goals but they can also produce perverse outcomes that frustrate policy objectives. Incentives can be financial and non-financial. For instance, Confucius emphasized the role of trust, without which “we cannot stand”. Trust creates duty, which is a clear non-financial behavioural incentive. O’Neill (2002) elaborated this theme in the context of individual and group behaviour. “Each of us and every profession and every institution need trust. We need it because we have to be able to rely on others acting as they say they will, and because we need others to accept that we will act as we say we will... .” Another way of looking at the problems associated with exchange between purchasers and providers is that contracts can never be complete. It is impossible to identify and include all possible behaviours in a contract, let alone legislate for their policing and management. As a consequence, trading between buyers and sellers has to be based on trust (Maynard & Bloor, 2003b). The importance of trust and duty should not be ignored when manipulating financial incentives. As demonstrated in the preceding discussion of market failures in health care, trust alone does not deliver efficiency or equity. It is necessary to sustain and develop trust by improving information systems, particularly in regard to measuring and managing outcomes. Improving the transparency of clinical and hospital performance may improve accountability and efficiency. This may facilitate change led by clinicians who are key decision-makers in health care. Financial incentives can be used marginally to complement trust. Trust and duty, together with incentives, are the primary determinants of behaviours and have to be carefully balanced as policy is developed.

Paying hospitals

In many public health care systems, a common form of paying hospitals was a block grant or a global budget. Typically such hospitals were reimbursed annually on the basis of what they received last year, plus allowances for expected inflation and some growth, with adjustments at the margin to reflect local media pressure and voting. This method of payment was used because, with financial discipline, it controlled expenditure inflation. This control has been evident in Scandinavian countries and the United Kingdom, but less evident in Italy and Spain. Global budgets may be a necessary but not sufficient means of achieving macroeconomic cost control, and they do not provide management with information about costs, activity volumes and outcomes. Consequently, global budgeting systems have had to invest in complementary performance management and more recently the development of prospective payment systems.

The hospital reimbursement system until recently paid hospitals in Germany on the basis of bed days used. This induced inefficient lengths of stay and consequently Germany, like England, has now followed a trend begun in the United States over 20 years ago of payment on the basis of diagnostic related group (DRG). In England the system is called “payment by results” (PbR).

Like the DRG systems elsewhere in the world, PbR is a system of payment for activity in relation to a schedule of diagnostic tariffs.

The incentive to increase activity created by DRGs can be constrained in a number of ways. The two-part tariff is a method used by public utilities (such as the water, electricity and gas industries) to manage demand. The English use this mechanism in the NHS. Elective activity is unconstrained as the government wishes to reduce waiting times. However, for emergency activity the full tariff is paid only for last year's volume of activity, plus 3%. Once activity exceeds this level, the tariff payments fall to 50%. The objective of this mechanism is to control demand for hospital emergency care and provide an incentive for the creation of alternative care facilities in the community.

Usually DRG tariffs are set in relation to the average cost of the procedure across some sample (or the total population) of hospitals. The sample from which the tariff is calculated varies from country to country and is sometimes representative of the population of hospitals and sometimes representative of the cheaper (and hopefully more efficient) hospitals. In England the tariff set for some procedures is not the overall average but that for particular producer groups. For example, where day case surgery is prevalent, the tariff is the average for day surgery activity, thereby giving producers an incentive for such techniques. This trend towards normative pricing is growing and will be discussed further below.

Another way in which tariffs can be manipulated to achieve service delivery goals is discrimination to encourage concentration of services when there are economies of scale. The literature on volume outcome relationships is extensive but of uneven quality. For some procedures there appears to be evidence of that concentration of activity in specialized locations gives superior patient outcomes measured in mortality rates (this literature does not include patient-reported outcome measurement), e.g. surgery for pancreatic cancer, oesophageal cancer, abdominal aortic aneurysms and paediatric cardiac problems (see Table 3).

Systematic reviews of this literature emphasise the methodological problems and the magnitude of volume–outcome effects varies significantly depending on whether administrative or clinical data are used for risk adjustment (Centre for Reviews and Dissemination, 1997; Halm et al., 2002). Those who have reviewed this literature are generally very critical of the methods used; Halm et al. (2002) found only 2 out of 135 evaluated the appropriateness of patient selection. The size of volume-outcome relationships in heart surgery, cancer surgery and orthopaedics appear to be small. Given the uncertainty about these relationships, discriminatory pricing to induce centralization has to be designed with care and moderation. These are not the characteristics of reforms suggested by the English government, which include tariff-induced centralization in cardiology, respiratory, orthopaedics, neurosciences and specialized children's services (Gainsbury, 2007).

Table 3: Economics of scale: the relationship between volume and outcome

Median absolute differences in mortality rate for high versus low volume (range) (%)		
High		
Pancreatic cancer surgery	13.0	3.0 to 17.0
Oesophageal cancer surgery	12.0	11.0 to 13.9
Paediatric cardiac surgery	11.0	2.3 to 15.5
AIDS	9.3	3.7 to 20.1
Surgery to repair ruptured abdominal aortic aneurysm	7.9	1.5 to 18.7
Gastric cancer surgery	6.5	4.0 to 7.1
Surgery for ruptured cerebral aneurysm	5.8 & 9.0	
Low		
Total knee replacement	0.1	0.0 to 1.4
Open prostatectomy	0.2	
CABG	0.2	
Coronary angioplasty	0.2	
Lung cancer surgery	1.9	
Myocardial infarction	2.3	

Notes: Outcome is death, in hospital or at 30 days. Volume is number of cases.
Source: Halm et al.

The focus on volume-outcome may lead to the relative neglect of volume–cost and volume–access considerations. Taking the limited volume-outcome data as the basis for centralization ignores volume-cost issues. The evidence from mostly American studies indicates that the average cost curve (with volume proxied by the number of beds) falls to around 600 beds and then rises. Thus any drive to reduce mortality outcomes by increasing volume and hospital size may drive up unit costs, and create difficult outcome-cost trade-offs for policy-makers. Furthermore, concentration of services on the basis of volume-outcome evidence will result in fewer units providing higher volumes of care. This will affect access for patients, especially their time costs. Increased time costs may differentially affect the utilization of care, particularly of the relatively disadvantaged that may have a greater incidence of conditions. The limited evidence base indicates that such time costs may affect the use of drugs and alcohol services and screening, but less so services for severe conditions such as cancer (Centre for Reviews and Dissemination, 1997).

The trend towards normative pricing is a product of evidence of growing international frustration with DRGs. Hospital tariffs based on average cost alone do not induce radical changes in efficiency. While there is financial pressure on those with above average costs to be more economical, those hospitals with below average costs initially tend to drift towards the mean i.e. DRGs reduce dispersion around the mean but may not shift it to a more economical level. DRGs create short-term effects on length of stay, where reductions may release some funding. Because of these tendencies, policy-makers are looking for complementary methods to improve efficiency. Nonetheless, such changes have both to be evidence-based and evaluated with care. Incentives can improve and worsen efficiency!

Paying doctors

The methods by which hospitals are remunerated have predictable effects on the behaviour of managers. If hospitals are given further incentives to improve efficiency, these tariffs may need to be supplemented with incentives bearing directly on the principle clinical decision-makers, i.e. specialists. Incentives at the hospital and specialist level should ideally be compatible and

complementary. JC Robinson (1999) noted: “There are many mechanisms for paying doctors; some are good and some are bad. The three worst are fee-for-service, capitation and salary”. The clear implication is that each of the principal methods of remunerating doctors has benefits and costs. With fee-for-service (FFS) the doctor receives a fee for each intervention in the payment schedule. The list of interventions may be long, as in Germany, or more selective, as in Britain. The appropriateness of incentivizing particular interventions will depend on evidence of their cost–effectiveness. Sadly, linkage of FFS payments to cost–effectiveness is less than complete. Without caps on expenditure, FFS payment systems can create expenditure inflation. The Germans cap their doctor pay fund with reductions in individual tariff payments if volumes increase excessively. In the United States the “managed care” approach to controlling expenditure led to some insurers moving away from FFS to capitation and salary systems as means of controlling inflation.

A salary reward system pays doctors to provide a certain minimum amount of time. While the FFS system informs decision-makers about the quantity and variation in services delivered, a salary system provides no information about cost, quantity or quality. However, given knowledge of doctor employment and a clear salary structure, it does offer expenditure control. Capitation systems of payment reward doctors for being available to treat patients on their list. Payments may be age-related, with more pay for potentially high users such as young children and the elderly. Like a salary system, a capitation system offers expenditure control but no information about cost, quantity and quality of care delivered.

The attributes of the three payment systems for doctors are summarized in Table 4. The choice of payment system depends on the goals that policy-makers are seeking to incentivize. The importance of Robinson’s remarks is that typically payment systems will be mixed or blended, and policy makers will use combinations of the three payment mechanisms to pursue both greater microeconomic efficiency and expenditure control.

Table 4: Attributes of payment systems for doctors

Type of pay	Incentive effects				
	Increase activity	Decrease activity	Shift costs	Target the poor	Control cost
Fee-for-service	yes	no	no	maybe	no
Salary	no	yes	yes	no	yes
Capitation	no	yes	yes	no	yes

Reinforcing the purchaser role: the development of normative incentives

Before describing and appraising the development of normative incentives in the two major centres of such reform (England and the United States), some general issues have to be noted. In discussion of P4P, there is scope for both non-payment for performance and incentives that penalize rather than reward. These are sometimes referred to as “reputational” incentives – rewards and penalties that may be marginal in financial terms but nonetheless provoke

significant behavioural reactions that improve efficiency. Non-reimbursement for failure may have a more powerful effect on performance than positive rewards. Indeed, as discussed in economic theory, a small financial loss may have a larger behavioural effect than a larger financial gain (Kahneman & Tversky, 1979). This approach, based on prospect theory, is being increasingly considered by health care reformers. With both rewards and penalties, the emphasis is on the experimental use of small incentives. The emerging evidence is that this approach, at the margin, may be effective at eliciting significant changes, in part probably as a result of the effects not just on revenue but also, via publicity, on public reputation.

Normative incentives for hospitals

American hospitals are undergoing radical changes in funding and incentives that are gradually being emulated in Europe (e.g., the NHS North West Strategic Health Authority's adaptation of the Premier approach; see below). England has also invested in the performance management of the NHS, sharing with the United States frustration with existing performance and perverse incentives.

For over 20 years there has been an increasing interest in the United States in the improvement of process management. This work on "total quality management" (TQM) was developed by the Joint Commission on Accreditation of Health Care Organizations (JCAHO), which sought to use both process measures and also some outcome measures, such as mortality. This work in the last decade has also sought to link diagnosis to treatment and ensure this is based on evidence. To encourage transparency, HEDIS has been deployed by insurers and large employers (e.g. General Motors) to collect performance data from health plans and facilitate comparison. In 2000 a group of private and public purchasers formed the Leapfrog Group which cumulatively adopted structure and process measures to improve the quality of health care delivery (Galvin et al., 2005). From this initial focus on process measures, in recent years there has been greater attention paid to crude outcome measures such mortality and postoperative complications. The patient safety movement, encouraged by the Institute of Medicine report (Kohn, Corrigan, Donaldson, 1999) also attracted greater attention to the mitigation of medical errors.

CMS-Premier

More recently the Centers for Medicare and Medicaid (CMS) and the Premier Hospital Quality Incentive Demonstration (HQID) have begun to create incentives for change in ten process measures. This programme is a public-private collaboration which brings together CMS, the federal agency responsible for Medicare, which is responsible for the healthcare of 40 million elderly Americans, and Premier Inc., which is an alliance of 1700 non-profit hospitals and health systems. CMS-Premier HQID covers over 260 hospitals in 37 states. Its primary focuses are acute myocardial infarction (AMI), heart failure, pneumonia, coronary artery bypass graft (CABG) and hip and knee replacements. Participating hospitals have to address all five clinical areas, pay fees for relevant software and staff the collection of data. The majority of the 34 quality measures used are process indicators, with seven basic outcome measures, e.g., mortality after AMI and CABG, postoperative complications and readmissions after discharge for hip and knee replacement.

Rewards for good performance are funded by Medicare. Hospital in the top quintile within each of the five categories may be rewarded. A hospital in the top decile in any one area receives a bonus of 2% of Medicare DRG payments in that clinical area. Performance in the second decile generates a 1% reward. Annual total payments in each of the first two years (2004, 2005)

exceeded \$8 million paid to over 100 hospitals. At the end of the initial 3-year period, penalties of 1% of revenues will be levied for poor performance between the eighth and ninth deciles, and 2% for those between the ninth and tenth deciles. The delayed penalties were organized to encourage performance improvement. In the first two years, CMS estimate that quality performance across the clinical areas improved by 11.8%, implying that hospitals were performing with closer adherence to clinical guidelines. The gap between best and worst performers is closing. These results have led to an extension of the programme for a second period of three years.

The obvious reservation about these findings is the absence of controls. Lindenauer and others (2007) showed higher levels of achievement of process targets in experimental hospitals that reported their performance and used modest P4P incentives, compared to hospitals that may only report their data. Another unanswered question is how the gains in CMS-Premier performance compare with the costs of the programme. The CMS also publish activity data, which provides another incentive for providers to improve their performance. All hospitals carrying out acute care for Medicare patients may be penalized by 0.4% of Medicare fees if they fail to report process quality data for 10 clinical measures. This reflects Anglo-American insistence on the disclosure of performance data, on the rationale that taxpayers deserve to be informed about performance, and such transparency facilitates accountability. Transparency is seen as an essential component of contracting between purchaser and provider; it is unusual not to see such transparency in Estonia. The downside of such openness is the difficulty of knowing which indicators to select and the transaction costs of evaluating system performance. These costs are clearly not inconsiderable and they have to be weighted carefully against the benefits. A related cost of transparency is the management of the information when it enters the public domain. Careful presentation of the process and outcomes data is necessary to avoid public confusion and political fall out.

Epstein (2007) and Lindenauer et al. (2007) referred to modest but significant impact in their discussion of the CMS-Premier results. The effects were for American hospitals already publicly reporting their results so it is possible that these institutions are not representative of the entire hospital population since such reporting is not universal.

Payment for non-performance

A recent innovation by CMS is non-payment for performance. This innovation will be introduced from October 2008 and is designed to reduce medical errors in hospitals by refusing reimbursement of conditions not present at the time of admission. This could significantly affect hospital revenue. For instance, if a patient is admitted with pneumonia and contracts bed sores or a urinary tract infection, the hospital would be paid for “pneumonia with complications” (currently \$6253). From October 2008, the hospital will be paid \$3705 for “simple pneumonia” (Rosenthal, 2007). The transaction costs of this system may be considerable and the incentives to game may increase. The volume of errors (non-performance) in US Medicare in 2006 is shown in Table 5.

Table 5: Non-payment for performance

	US Medicare: medical errors in 2006
1	Pressure ulcers (322 946 cases)
2	Catheter associated urinary tract infections (11 780 cases)
3	Falls from bed (2591 cases)
4	Objects left in patients after surgery (764 cases)

Source: Rosenthal, 2007

Patient Reported Outcome Measures (PROMs)

There is some evidence on the use of PROMs. In Dartmouth and Cleveland, for example, SF36 is being used for some aspects of cardiology. Overall the United States remains focused largely on process improvement and the interventions are local and fragmented. However, there is recognition of the challenges of the quality agenda, and some small signs that it is progressing to complement process with outcome measurement and management.

Most of the efforts to improve process and outcome quality in the United Kingdom have been in England. The other three countries of the United Kingdom (Northern Ireland, Scotland and Wales) watch the English experiments with interest, but are not obliged to emulate them since their own NHS systems are independently managed. For over 20 years there has been increasing emphasis on performance management in the NHS. This was reinforced by Thatcher's mantra of "value for money". Systematic medical auditing was propelled onto the political agenda by the National Confidential Enquiry into Perioperative Deaths (NCEPOD) (Buck, Devlin, Lunn, 1987). This innovative work did not ensure professional participation and a decade later, despite Thatcher's efforts to mandate medical auditing in the NHS, as many as a third of surgeons and anaesthetists in some regions were still not contributing their data (Warden, 1998). Mandatory inclusion of all practitioners' data is essential.

A series of medical disasters involving individual practitioners in the late 1990s and early 2000s, together with increased concern about medical errors demanded improved measurement and management of clinical performance (Vincent, Ennis, Audley, 2004; Sari et al., 2007). The medical disasters included gynaecological surgeons who rendered hysterectomy patients incontinent in large numbers, a paediatric cardiac surgeon at whose hand nearly 30 children died and a GP who appears to have poisoned over 200 elderly people with fatal morphine injections. As these problems were recognized, the Blair administration sought to improve the NHS by structural reorganization and after a decision to sharply increase funding, enhanced regulation and performance management was even more important. The result was a significant investment in process performance indicators with a particular focus on driving down waiting time.

The current regulatory agencies include the Healthcare Commission (HCC), which focuses on a portfolio of weighted process measures used to grade the clinical and financial performance of all purchasers and providers (public and private) trading in the NHS. In addition, the agency Monitor to some degree duplicates the HCC controls, but regulates a select group of Foundation Trust hospitals, which are generally the better performing ones. The Royal College of Physicians complements this with government sanctioned performance in relation to stroke care and the delivery of thrombolytics. The Society of Cardiac Surgeons has created a publicly accessible data base showing the relative performance of its members. This is accessible via the HCC website and gives risk-adjusted mortality rates for surgeons working in the NHS

(<http://heartsurgery.healthcarecommission.org>) These data not only inform practitioners of their relative success but also inform patients deciding where to have their treatment.

For the reasons discussed earlier the NHS in England is now increasingly focused on the application of PROMs to routine clinical practice after decades of using specific and generic measures in clinical trials. The English Department of Health announced the use of PROMs in the NHS from April 2009 (Department of Health, 2007). This will cover hip and knee replacement, hernia repair and varicose vein procedures. For each condition, specific and a generic quality of life measure is advised (see Table 6).

Table 6: Measuring patient outcomes in the English NHS

Procedure	Condition-specific	Generic
Primary unilateral hip replacement	Oxford Hip Score	EQ5D
Primary unilateral knee replacement	Oxford Hip Score	EQ5D
Groin hernia repair	None	EQ5D
Varicose vein procedures	Aberdeen Varicose Vein Questionnaire	EQ5D
Plus a standard set of patient-specific questions in all cases		

Source: DH Operating Framework

There are several inherent problems with the use of PROMs, starting with cost. English estimates put the cost of data collection before and after treatment and its analysis at around £10 per patient: approximately equal to the cost of a full blood test. Another challenge is response rates. Ideally, they need to be around 80–80, i.e. 80% completed PROMs before and after the procedure. BUPA created an incentive of performance-related pay for managers who got completed patient returns. An alternative might be rewards for patients. Such incentives inevitably affect the cost of a PROMs programme. The final potential problem is case mix adjustment. Adjustment for the age and gender of respondents will be routine, but outcomes may reflect complex co-morbidities as much as the success of the procedure.

These challenges in the design and implementation of PROMs should ensure a gradual approach, with careful evaluation to maximize the efficiency of this outcome measurement programme. Given the novelty of this work, the need for cautious development with evaluation is as urgent as it is unlikely if politicians become impatient with the low process productivity of health care and seek to divert criticism by focusing public attention on PROMs. The case for cautious innovation with PROMs is clear. Its approval may be very useful to management in areas such as elective surgical procedures, but the obvious risk is that politicians may be too ambitious and resist the need to innovate with evaluation.

Normative incentives for doctors

A good example of normative incentives for doctors is the 2004 United Kingdom contract for general practitioners, which introduced the quality outcomes framework (QOF), largely an attempt to increase delivery of chronic care. The rewards are generous and paid to practices, i.e., groups of GPs, in an effort to ensure peer management and high uptake. The fee-for-service system was initially based on “light touch” performance management of ten clinical categories.

Performance was related to prevalence and a target for performance, and full achievement of the items now earns 1000 points at £125 per point (Maynard & Bloor, 2003b; Doran et al., 2006). This reform was designed to improve the delivery of care to the chronically ill. Its intention was excellent but its execution was less than perfect. The reform offers the opportunity for other reformers to learn and improve the design and delivery of P4P policy.

In the United Kingdom, sadly, no baseline data was collected to facilitate a before-and-after evaluation of the reform. The reform was very costly (£1 billion), and it contributed to an average increase in general practitioner pay of 23%. Achievement of QOF targets anticipated a 75% performance in the first year, but GPs delivered over 90% of the target, with incomplete data suggesting improving performance before 2004. The opportunity cost of the reform, in particular identifying what services were reduced as practices pursued the QOF targets, is unknown. Furthermore, as is common with fee-for-service policies, there has been criticism of the evidence used to select the interventions (Cookson & Fleetcroft, 2006). It is debatable whether the items included maximize population health improvement. Despite such caveats, American commentators have expressed enthusiastic support of the United Kingdom's QOF reform (Epstein, 2006). Recent research shows evidence of gaming (Gravelle, Sutton, Ma, 2007). Despite these problems, many of which could have been anticipated with more careful policy design, the GP-QOF in the United Kingdom may have created a major, largely beneficial change in health care delivery.

New systems of incentives for hospital doctors have also been used in the United Kingdom and the United States. During discussion about the new NHS specialist contract, it was proposed that using a fee-for-service system might reduce the dispersion in consultant activity (see Figs 2 and 3 above, and discussion). The basis of this discussion was the need to increase the activity and productivity of consultants and obviate the need for increased medical training. This proposal was not adopted, although strongly supported by economists. In Boston, private sector hospital groups are innovating reforms based on prospect theory. This involves setting specialists' performance criteria, evaluating performance and reducing specialists' incomes by small amounts when they fail to perform adequately. The evidence to support this approach is limited but indicative (Rizzo & Zeckhauser, 2003). Such reforms in Medicare have proved impossible. Physicians in Medicare are paid in relation to a relative value scale. Over 40 years, this has become inefficient and difficult to change (Newhouse, 2007).

There are clear links between using penalties and rewards at hospital level and specialist levels. Their advantage is their relative cost-effectiveness, singly and in combination. A priori, the case for using them as complements appears to be strong, particularly as hospital incentives systems seem to be achieving only modest performance improvements. The lessons to be learnt from these innovative approaches to physician contract reform may be summarized as follows.

- Be clear about the goal of the policy and the evidence base of the instruments used, e.g. do they provide population health gain in the most cost effective manner?
- Evaluate the policy carefully by collecting before and after data.
- Be prepared for gaming and recognize that its identification and policing may impose significant transaction costs on the health care system.
- Do not be dissuaded from reform because it is complex and the potential for error is always considerable. The cost and benefits of change are considerable but the costs of lethargy and the status quo are unacceptably high in relation to damage to patients and tax payers!

Conclusions

Internationally, there is frustration with the rigid and conservative nature of health care delivery systems. Decision-makers, managers and clinicians tend to ignore the evidence base, exhibit large variations in clinical care, fail to deliver cost-effective interventions and focus unduly on reform of structure and process, with too little regard for patient outcomes. Intolerance of this inefficiency is leading to more radical P4P reforms.

What lessons can be learnt from the P4P literature, especially for purchaser agencies such as the EHIF? The first policy lesson from the literature is that there is much uncertainty about the effects of health care investment on the health of the population. Much of health care lacks an evidence base about clinical effectiveness, let alone cost-effectiveness. As a consequence, there is a risk that unless health care investment is carefully targeted the health gain may be slight due to diminishing returns.

The structure of the Estonian health care system is well established. Any reform of it and the processes by which it commissions care should be undertaken carefully and with an emphasis on exploiting available evidence and contributing to knowledge of how trading incentives affect the efficiency of health care delivery. Internationally, policy-making is often based on good intentions, but “faith-based” reforms consume scarce resources, with significant opportunity costs in terms of reduced patient care. The reform process should begin with policy-makers debating and answering the fundamental questions set out in Table 7. Understanding of these questions and answers is an essential part of reaching a consensus between EHIF and hospital managers about how their system performs.

Table 7: Reform design: prior questions

1	What are the objectives of the health care system? What ordering or weight do these objectives have, and how do they change over time? An analysis of policy statements by government and EHIF can produce answers to these questions.
2	Who is really responsible for control of the system(s): who controls resource use at the boundaries of care, for example, between primary and hospital care? Who controls movements across these boundaries? What criteria determine policy-making at these boundaries?
3	What incentives (monetary and non-monetary) are there to achieve efficiency for individual managers (clinical and non-clinical) and for institutions? Why do decision-makers at the boundaries behave as they do?
4	Who rations what and how? What criteria are used by decision-makers to allocate resources? Are the rationing criteria consistent with policy objectives set out in 1 above?
5	Who in effect decides to allocate resources, and what are the investment criteria?
6	What are the major unresolved problems of the system?

Source: Maynard, 2005

The next step in collaborative EHIF-hospital reform is agreement on the nature of the health care failures and their prioritization. The continuing failure to link the reform of structure and process to their effect on patient outcomes ensures that clinical practice variations are sustained despite research and policy literature going back decades. The increasing frustration of policy-makers with practice variations has led to an increasing focus on P4P incentives, led particularly by the

Americans and the English. This work continues to be concentrated on process reform. Reducing variation through incentives for greater consistency in clinical behaviour is to be welcomed provided it is linked clearly to the improvement of outcomes for patients.

The current wave of P4P process initiatives, particularly in the United States, are using wider ranges of quite simple incentives. Typically they are evaluated in terms of process effect and little attempt appears to be made to determine their overall cost–effectiveness, let alone that of their individual components. As the transaction costs of P4P rise, the need for such data will become more apparent. P4P investments will have to be prioritized and to do this a cost–effectiveness evidence base, built on controlled evaluations, will be essential.

The emerging P4P process evidence base is demonstrating effect. However, the initiatives are becoming increasingly ambitious e.g., Medicare’s 2008 non-payment for medical errors (Rosenthal, 2007). Such initiatives may not only have high transaction costs, they may also induce increased gaming. The incentives literature is replete with examples of P4P mechanisms creating opportunities for “creative” management (e.g., DRG creep, where patients with multiple morbidities are evaluated by hospitals to maximize tariff revenue). Such ubiquitous behavioural responses have to be policed, further inflating transaction costs. However, such problems have to be seen in light of the potential gains of knowing more about hospital activity and how this basic information can improve management and patient care.

An interesting issue is that of the size of incentive required to induce change in doctors and providers. Prospect theory and the idea of reputational incentives imply that small negative incentives (income losses) may produce more change than larger positive incentives (bonuses). The Medicare–CMS hospital incentives are small and produce change and there is some evidence that small negative incentives may affect doctors, too (Rizzo & Zeckhauser, 2003).

The enthusiasm of P4P policy advocates should be welcomed, provided such social experiments involve marginal and well-articulated policy changes that are carefully evaluated with robust comparators. Policy reform and its evaluation must be promoted on the basis of evidence of cost–effectiveness rather than faith. An essential part of this work is the use of PROM to demonstrate that as processes are improved and clinical practice variations are reduced, there is an assurance that patients are “getting better”. Without PROMs, transparency and accountability will be difficult to achieve. Pro-active commissioning by EHIF, focusing on quantity and cost will always risk adversely affecting the quality of patient care in terms of PROMs. Without PROMs, purchasers operate in the dark, and there is little consumer protection for the taxpayer and the patient. With PROMs, process data can be complemented and improved clinical and non-clinical management enabled. Cautious, incremental investment in outcome measurement may cast light on whether patients’ physical and psychological functioning improve.

The policy process involves discrete steps with resource consequences at each stage. These are set out in Table 8.

Table 8: The reform policy process

1	Clearly specify the process and/or outcome objectives of the reform.
2	Create provider acceptance of the need for reform by collaborating with them throughout the process.
3	Invest in appropriate design and implementation of process and outcomes information systems. This will involve identification and agreement with providers of a minimum data set and sharing of the data. Investment in medical coding staff and auditing will also be essential.
4	P4P design: ensure congruence between the objectives and the policy design. Consider the comparison of budget neutral and gain/loss incentive systems. Cost-estimate these systems, particularly the management costs. Ensure timely data collection and feedback of results to providers and practitioners.
5	Manage implementation, monitoring and evaluation with providers, government and the public.
6	Identify and agree indicators of success in relation to the objectives of the reform process.

Adapted from Lindenauer et al., 2007 and Schneider, 2007

Campbell (1969) emphasized nearly 40 years ago that all reform is social experimentation. Health care reform imposes costs and benefits on taxpayers and patients. These should be as carefully evaluated as a new pharmaceutical when it is brought to market. Health care reforms and pharmaceuticals can seriously damage the welfare of society if not evaluated rigorously.

Finally, these conclusions should induce excitement that incentive redesign has reached the top of the health reform agenda but also a sense of caution when reforming health care with P4P. Incentives – monetary and non-monetary – are powerful in their effects on behaviour and effort has to be made to improve performance and avoid perverse effects that damage the financial and health interests of patients and taxpayers. Reformers in affluent countries in recent decades have failed to make significant progress in mitigating the problems identified in the research literature. This is often a product of confusing action and evidence-based policy-making. Well-intentioned, faith-based policy reform has dominated international health care reform with consequent frustration and resource wasting. As Campbell noted, “there is safety behind the veil of ignorance”. The cost of political safety for patients and taxpayers is very high! As the then President of the Royal College of Surgeons of England said of the Thatcher government’s reforms, “instead of ready, take aim and fire, the Government made ready, fired and then took aim” (personal communication to the author).

The challenge for EHIF and health care providers is to learn the lessons of the international literature when developing P4P. These lessons are quite obvious and require collaboration in the careful design of any reforms, with robust piloting and evaluation. Some may be intimidated by these challenges and prefer the quiet life of accepting existing inefficiencies in health care delivery that penalize Estonian patients and taxpayers. The costs of P4P reform are likely to be considerable, but the costs of such inertia are even greater.

References

- Atkinson Sir Tony (2005). *Atkinson Report on measuring government output and productivity*. London, Office for National Statistics.
- Bernstein SJ et al. (1993). The appropriateness of the use of cardiovascular procedures. British versus U.S. perspectives. *International Journal of Technology Assessment in Health Care*, 9(1):3–10.
- Bloor K, Maynard A (2002). Consultants: managing them means measuring them. *Health Service Journal*, 112(5836):10–11.
- Bloor K, Maynard A (2006). Consultant clinical activity is key to improving productivity. *Health Service Journal*, 116(6013):18–19.
- BMJ (2007). *BMJ Clinical Evidence Handbook*. London, BMJ Publishing.
- Brownlee S (2007). Overdose. *The Atlantic Monthly*, December (<http://www.theatlantic.com/doc/2007/12/health-care>, accessed 19 March 2008).
- Buck N, Devlin HB, Lunn JN (1987). *Report of the National Confidential Enquiry into Perioperative Deaths*. London, The Nuffield Provincial Hospital Trusts and King Edward's Hospital Fund.
- Campbell, DT (1969). Reforms as experiments. *American Psychologist*, 24(4):409–429.
- Centre for Reviews and Dissemination, 1997. Hospital volume and outcomes, cost and patient access. *Effective Health Care*, 2(8) (<http://www.york.ac.uk/inst/crd/pdf/ehc28.pdf>, accessed 19 March 2008).
- Cochrane AL (1972). *Effectiveness and efficiency*. London, The Nuffield Provincial Hospitals Trust.
- Cookson R, Fleetcroft R (2006). Do incentive payments in the new NHS contract for primary care reflect likely population health gain? *Journal of Health Services Research and Policy*, 11(1):27–31.
- Department of Health (2007). *The NHS in England: The operating framework for 2008/9*, London, Department of Health.
- Department of Health and Social Security (1976). *Priorities in health and personal social services*. London, Her Majesty's Stationery Office.
- Donabedian A (1966). Evaluating the quality of medical care. *Milbank Memorial Fund Quarterly*, 44:166–206.
- DH Operating Framework (2007). *Guidance on the routine collection of patient-reported outcome measures*. London, Department of Health.

Doran et al. (2006). Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine*, 355(4):375–384.

Epstein AM (2006). Paying for performance in the United States and abroad, *New England Journal of Medicine*, 355(4):406–408.

Epstein, AM (2007). Payment for performance at the tipping point. *New England Journal of Medicine*, 356 (5):515–517.

Fisher ES (2003). Medical care: is more always better? *New England Journal of Medicine*, 349(17):1665–1667.

Galvin RS et al. (2005). Has the Leapfrog group had an impact on the health care market? *Health Affairs*, 24(1):228–233.

Gainsbury S (2007). District general hospitals face heavy specialist service losses. *Health Service Journal*, 5 (8 November).

Gravelle H, Sutton M, Ma A (2007). Doctor behaviour under a pay for performance contract: Evidence from the quality and outcomes framework. York, Centre for Health Economics (CHE Research Paper no. 28).

Grol R, Berwick DM, Wensing M (2008). On the trail of quality and safety in health care. *British Medical Journal*, 336:74–76.

Halm EA, Lee C, Chassin MR (2002). Is volume related to outcome in health care? A systematic review and methodological critique of the literature. *Annals of Internal Medicine*, 137(6):511–520.

Kahneman D, Tversky A (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47:263–292.

Kerr EA et al. (2004). Profiling the quality of care in twelve communities: results from the CQI study. *Health Affairs*, 23(3):247–256.

Kind P, Williams A (2004). Measuring success in health care: the time has come to do it properly, *Health Policy Matters* 9, (<http://www.york.ac.uk/healthsciences/pubs/HPM9final.pdf>, accessed 15 March 2008).

Kohn LT, Corrigan JM, Donaldson MS, eds. (1999). *To err is human: Building a safer health care system*. Washington, Institute of Medicine, Academy of Medical Sciences.

Lindenauer PK et al. (2007). Public reporting and pay for performance in hospital quality improvement. *New England Journal of Medicine*, 356(5):486–498.

Maynard A (1997). Evidence-based medicine: an incomplete method for informing treatment choices. *Lancet*, 349(9045):126–128.

Maynard A, ed. (2005). *The public-private mix for health: plus ça change, plus c'est la même chose*. London, Radcliffe Publishers for the Nuffield Trust.

- Maynard A, Bloor K (2003a). Trust and management in the medical market place. *Journal of the Royal Society of Medicine*, 96:532–539.
- Maynard A, Bloor K (2003b). *Do those who pay the piper call the tune?* York, University of York.
- McPherson K et al. (1982). Small-area variations in the use of common surgical procedures: an international comparison of New England, England and Norway. *New England Journal of Medicine*, 307(21):1310–1314.
- Newhouse JP (2007). Medicare spending on physicians: no easy fix in sight. *New England Journal of Medicine*, 356(18):1883–1884.
- NHS Institute for Innovation and Improvement (2006). *Delivering quality and value: focus on productivity and efficiency*. London, Department of Health.
- Nightingale F (1863). *Some Notes on Hospitals*, 3rd edition. London, Longman, Roberts and Green.
- Nolte E, McKee M (2004). Does health care save lives? Avoidable mortality revisited. London, Nuffield Trust for Research and Policy Studies in Health Services.
- O'Neill, O (2002). *A question of trust*. Cambridge, Cambridge University Press (BBC Reith Lectures).
- Patrick DL, Erickson P (1993). Health status and health policy: Quality of life in health care evaluation and resource allocation. Oxford, New York, Oxford University Press.
- Pittet D et al. (2000). Effectiveness of a hospital-wide programme to improve compliance with hand hygiene. *Lancet*, 356:1307–1312.
- Pittet D, Boyce JM (2001). Hand hygiene and patient care: pursuing the Semmelweis legacy. *Lancet Infectious Diseases*, April:9–20.
- Rizzo JA, Zeckhauser RJ (2003). Reference incomes, loss aversion and physician behaviour. *Review of Economics and Statistics*, 85(4):909–922.
- Robinson JC (1999). Blended payment methods in physician organizations under managed care. *Journal of the American Medical Association*, 282(13):1258–1263.
- Rosenthal MB (2007). Non payment for performance? Medicare's new reimbursement rule. *New England Journal of Medicine*, 357(16):1573–1575.
- Sari AB et al. (2007). Extent, nature and consequences of adverse events: results of a retrospective case note review in a large NHS hospital. *Quality & Safety in Health Care*, 16:434–439.
- Select Committee on Health (2008). *The National Institute of Health and Clinical Excellence (NICE) Report on an investigation by the Select Committee on Health*, Westminster, House of Commons.

Stewart AL, Ware JE, eds. (1992). *Measuring functioning and well-being*. Durham, NC, Duke University Press.

Vincent C, Ennis M, Audley RJ (2004). Analysis of clinical incidents: a window on the system not a search for root causes. *Quality and Safety in Health Care*, 13(4):242–243.

Vallance-Owen A, Cubbin S (2002). Monitoring national clinical outcomes: a challenging programme. *British Journal of Health Care Management*, 8(11), 412–417.

Wanless D et al. (2007) *Our future health secured? A review of NHS funding and performance*. London, King's Fund.

Warden J (1998). NHS doctors face compulsory audit. *British Medical Journal*, 316(1851)

Wennberg JE, Gittelsohn A (1973). Variations in medical care among small areas. *Science*, 246(4):100–111.

Wennberg JE, Freeman JL, Culp WJ (1987). Are hospital services rationed in New Haven or over-utilized in Boston? *Lancet*, 1(8543):1185–1189.

Wennberg JE et al. (1989). Hospital use and mortality among Medicare beneficiaries in Boston and New Haven. *New England Journal of Medicine*, 321:1168–1173.