PART V

Health policy and performance measurement

5.1 *Targets and performance measurement*

PETER C. SMITH, REINHARD BUSSE

Introduction

Targets are a tool designed to improve health and health system performance. They can facilitate the achievement of health policy by expressing a clear commitment to achieve specified results in a defined time period and facilitating the monitoring of progress towards the achievement of broader goals and objectives. They may be quantitative (e.g. x% increase in the immunization rate) or qualitative (e.g. introduction of national screening programme); based on health outcomes (e.g. reduction in mortality) or processes (e.g. reduction of waiting time). The introduction of the concept of targets into the health sector is often traced to the 1981 publication of WHO's *Health for All* strategy which presented targets as a tool with which to improve health policy (WHO Regional Office for Europe 2005).

Earlier chapters of this book discuss the manifest need for tools designed to improve performance and accountability. Thus it is not surprising that targets' role in health policy has grown and an increasing number of countries and/or regions now use them as tools to improve performance. Various mapping exercises have documented growing and sustained interest in health targets among governments and international organizations (Busse & Wismar 2002; Ritsatakis et al. 2000; van de Water & van Herten 1998). The 2005 update of the WHO European *Health for All* policies reported that forty-one of the (then) fifty-two Member States of the Region had either adopted or drafted policies which included health targets (WHO Regional Office for Europe 2005). Most recently, Wismar et al. (2008) offered many national and sub-national examples from Europe, primarily in population health. The Millennium Development Goals introduced important health targets at the international level.

A large body of literature has developed to provide increasing insights into the various dimensions of target setting and monitoring.

For example, there has been much discussion about the relative merits of goals that are process or outcome oriented. As explained below, we would argue that in reality this is a false dichotomy. Other debates have focused on the extent to which targets should set a general direction of travel or be detailed road maps, indicating every point along the way. This has been addressed by separating aspirational, managerial and technical targets that are ranked in terms of the extent to which they prescribe what should be achieved and how (van Herten & Gunning-Schepers 2000). Similarly, much has been written about the optimal characteristics of targets. At the risk of simplification, this literature has been reduced to a mnemonic, indicating that targets should be SMART – **s**pecific, **m**easurable, **a**chievable, **r**ealistic and **t**imed.

Rather than providing a systematic review of the issues surrounding the use of targets in the health sector, this chapter seeks to illustrate the general issues and to explore how targets contribute to improving health system performance. We use the specific example of the extensive English experience (possibly one of the most ambitious of such innovations to date) but also take account of experience in other European countries. The chapter begins with a brief history of targets in England. We then describe in some detail experience with the Public Service Agreement (PSA) targets introduced in 1998, under which targets assumed a much more central role. The chapter assesses the strengths and weaknesses of the PSA targets regime and concludes with the general lessons that can be learned from the English and European experiences.

Targets in the English health system

England has an extended history of targets in health and health care (Hunter 2002) but the first concerted attempt to introduce targets into English public health was the *Health of the Nation* strategy, launched in 1992 (Department of Health 1992). Owing a heavy debt to the WHO *Health for All* initiative, this was intended to encourage health authorities to focus on securing good health for their population. *Health of the Nation* can be seen as an attempt to set the public health agenda for local health authorities in the reformed NHS. Initially, five key areas were selected for action:

1. coronary heart disease and stroke
2. cancers

3. mental illness
4. HIV/AIDS and sexual health
5. accidents.

A small number of national targets were specified for each key area. For example, the targets for the first key area were:

- to reduce death rates for both coronary heart disease and stroke in people under 65 by at least 40% by the year 2000;
- to reduce the death rate for coronary heart disease in people aged 65-74 by at least 30% by the year 2000;
- to reduce the death rate for stroke in people aged 65-74 by at least 40% by the year 2000.

A careful independent evaluation of *Health of the Nation* in 1998 concluded that its: 'impact on policy documents peaked as early as 1993; and, by 1997, its impact on local policymaking was negligible' (Department of Health 1998). It found that health authorities felt that they had more pressing concerns than public health and therefore concentrated on operational issues, such as reducing waiting times and securing budgetary control. The evaluation concluded that the high-level national targets did not resonate with local decision-makers: 'National targets were a useful rallying point, but the encouragement to develop local targets would have been welcomed within the national framework as a reflection of local needs.' There was also seen to be a lack of incentives and institutional capacity for local managers.

Hunter (2002) summarizes the weaknesses of the *Health of the Nation* strategy under six broad headings.

1. Appeared to be a lack of leadership in the national government.
2. Policy failed to address the underlying social and structural determinants of health.
3. Targets were not always credible and were not formulated at a local level.
4. Poor communication of the strategy beyond the health system.
5. Strategy was not sustained.
6. Partnership between agencies was not encouraged.

The overarching theme was that the *Health of the Nation* strategy, and the associated targets, did not permeate the health system strongly enough to make a material difference.

The Labour government came to power in 1997 with a commitment to evidence-based policy; systematic priority setting; and explicit performance targets throughout the public services. A series of biennial spending reviews was implemented in 1998, setting three-year budgets in advance for each government department. Following the conclusion of the budgetary agreements, a set of PSAs with each department was announced. These were intended to signal priorities across the entire range of government activity and took the form of a series of specific objectives, expressed as a target in measurable form, that were expected to be achieved within a designated time frame. In common with other ministries, the Department of Health was set a series of PSA targets – for health and health care.

One distinctive feature of PSAs was the intention to focus on the outcomes of the public services rather than the operational activities of public service delivery. The PSA process signalled the government's determination to make the management of public services more transparent and to give departments clear statements of priorities. In the first round, the detail, specificity and measurability of the PSA targets were highly variable. However, over subsequent series of spending reviews the targets have become fewer and focused increasingly on outcomes.

An example: 2004 PSAs for the Department of Health

We illustrate the issues by describing the 2004 PSA targets which were based on four broad objectives.

1. Improve the health of the population. By 2010 increase life expectancy at birth in England to 78.6 years for men and to 82.5 years for women.
2. Improve health outcomes for people with long-term conditions.
3. Improve access to services, in particular waiting times.
4. Improve the patient and user experience.

The detailed targets associated with the objectives are given in Box 5.1.1; the four standards that must be maintained are shown at the bottom. These reflect targets secured through previous PSAs that must continue to be achieved. A set of even more detailed technical notes accompanies the targets, giving the context, data sources and measurement instruments. Box 5.1.2 gives an example, showing the technical note for the obesity target.

Box 5.1.1 Department of Health PSA Targets, 2004

Objective I: Improve the health of the population. By 2010 increase life expectancy at birth in England to 78.6 years for men and to 82.5 years for women.

1. Substantially reduce mortality rates by 2010:
   - from heart disease and stroke and related diseases by at least 40% in people under 75, with at least a 40% reduction in the inequalities gap between the fifth of areas with the worst health and deprivation indicators and the population as a whole;
   - from cancer by at least 20% in people under 75,with a reduction in the inequalities gap of at least 6% between the fifth of areas with the worst health and deprivation indicators and the population as a whole; and
   - from suicide and undetermined injury by at least 20%.
2. Reduce health inequalities by 10% by 2010 as measured by infant mortality and life expectancy at birth.
3. Tackle the underlying determinants of ill health and health inequalities by:
   - reducing adult smoking rates to 21% or less by 2010, with a reduction in prevalence among routine and manual groups to 26% or less;
   - halting the year-on-year rise in obesity among children under 11 by 2010 in the context of a broader strategy to tackle obesity in the population as a whole; and
   - reducing the under-18 conception rate by 50% by 2010 as part of a broader strategy to improve sexual health.

Objective II: Improve health outcomes for people with long-term conditions.

4. To improve health outcomes for people with long-term conditions by offering a personalized care plan for vulnerable people most at risk; and to reduce emergency bed days by 5% by 2008, through improved care in primary care and community settings for people with long-term conditions.

Box 5.1.1 cont'd

Objective III: Improve access to services.

5. To ensure that by 2008 no-one waits more than 18 weeks from GP referral to hospital treatment.
6. Increase the participation of problem drug users in drug treatment programmes by 100% by 2008 and increase year on year the proportion of users successfully sustaining or completing treatment programmes.

Objective IV: Improve the patient and user experience.

7. Secure sustained national improvements in NHS patient experience by 2008, as measured by independently validated surveys, ensuring that individuals are fully involved in decisions about their healthcare, including choice of provider.
8. Improve the quality of life and independence of vulnerable older people by supporting them to live in their own homes where possible by:

    - increasing the proportion of older people being supported to live in their own home by 1% annually in 2007 and 2008; and
    - increasing, by 2008, the proportion of those supported intensively to live at home to 34% of the total of those being supported at home or in residential care.

Standards

- A four hour maximum wait in Accident and Emergency from arrival to admission, transfer or discharge.
- Guaranteed access to a primary care professional within 24 hours and to a primary care doctor within 48 hours.
- Every hospital appointment booked for the convenience of the patient, making it easier for patients and their GPs to choose the hospital and consultant that best meets their needs.
- Improve life outcomes of adults and children with mental health problems, by ensuring that all patients who need them have access to crisis services and a comprehensive Child and Adolescent Mental Health Service.

Source: HM Treasury 2004

Box 5.1.2 Example of a PSA Technical Note – 2002 Joint Obesity Target for Department of Health (DH) and Department for Education and Skills (DfES)

PSA Target: Halting the year-on-year rise in obesity among children under eleven by 2010, in the context of a broader strategy to tackle obesity in the population as a whole.

Scope: children aged between two and ten years (inclusive) in England.

Obesity: prevalence of obesity as defined by the National BMI percentile classification (from the 1990 reference population from TJ Cole et al.) and measured through the Health Survey for England. Children above the 95th percentile of the 1990 reference curve are defined as obese.

Halt the year-on-year increase: obesity in two- to ten-year-olds rose, on average, by 0.8% per year between 1995 and 2002. Halting the increase would mean no significant change in prevalence between the two three-year periods 2005/06/07 and 2008/09/10.

Data source: Health Survey for England. We are also exploring with colleagues in DH and DfES the cost and feasibility of options for other sources of data in order to obtain more local level information.

Baseline year: due to the small sample size, the baseline will be the weighted average for the three-year period 2002/03/04.

Target year: by 31 December 2010, in practice this will mean 2010–2011 financial year.

Reporting: annually (aggregate trend data will be available every three years). The lag between the end of the collecting period and data being published is around twelve to fifteen months.

OGD contributions to PSA: delivery of this joint PSA target will be supported by a range of programmes including:

a) joint DfES and DCMS[1] PE, School Sport and Club Links project which seeks to increase the percentage of school children who

[1] Department for Culture, Media and Sport

Box 5.1.2 cont'd

spend a minimum of two hours each week on high quality PE and school sport within and beyond the curriculum;

b) joint DfES and DH National Healthy Schools Programme which seeks to promote a whole school approach to healthy living;

c) joint DfES and DH 'Food in Schools' programme which seeks to promote a whole school approach to a range of food issues.

Throughout the PSA regime, one of the Department of Health's central tasks has been to devise operational instruments that transmit the national PSA targets to the local level. To this end, the most important initiative was the development of a system of performance ratings for individual NHS organizations. Beginning in 2001, every organization (including local health authorities and NHS providers of care) was ranked annually on a four-point scale (zero to three stars) according to a series of about forty performance indicators. The indicators were intended directly to reflect the objectives of the NHS, as embodied in the national PSA targets (Department of Health 2001).

For each NHS organization, the star rating was produced by combining the indicators according to a complex algorithm. The most important determinant of an organization's rating was its performance against a set of about ten 'key indicators', which were then dominated by measures of various aspects of patient waiting times. This was augmented by a composite measure of performance based on the thirty or so subsidiary indicators, combined in the form of a balanced scorecard view of the organization. Clinical quality comprised only a small element of the calculation. In 2004 the health-care regulator took over responsibility for preparing the star ratings.

The most striking innovation associated with performance ratings was the introduction of very strong managerial incentives dependent on the level of attainment. Some commentators characterize this as a regime of terror (Bevan & Hood 2006b). The jobs of senior executives of poorly performing organizations came under severe threat and the performance indicators (especially the key targets) became a prime focus of managerial attention. Rewards for performing well included some element of increased organizational autonomy. For example, the best performers in the acute hospital sector became eligible to apply

for Foundation status which carries considerably greater autonomy from direct NHS control.

NHS managers have shown a mixed response to performance ratings. Many have criticized the system because of some of the apparently arbitrary ways in which the ratings are calculated and their sensitivity to small data fluctuations (Barker et al. 2004). However, some acknowledge that the system gives managers better focus and a real lever with which to affect organizational behaviour and clinical practice. Healthcare professionals have shown less ambiguous reactions and there is a widespread view that the ratings distort clinical priorities and undermine professional autonomy (Mannion et al. 2005). This is hardly surprising, as one of the aims of the national and local targets was precisely to challenge traditional clinical behaviour and to direct more attention to issues that had not always been a high priority e.g. waiting times.

There is no doubt that performance ratings have delivered major improvements in the aspects of NHS care targeted (Bevan & Hood 2006b). They have also secured marked progress towards some of the PSA targets. For example, very long waits for non-urgent inpatient treatment were a prime focus of the PSA regime and have been rapidly eliminated. Moreover, targeted aspects of English health care have improved markedly in comparison to Wales and Scotland, even though they have higher funding levels. These countries have not been subject to the PSA regime and have not implemented performance ratings (Hauck & Street 2007; Propper et al. 2008).

Less satisfactorily, the high level PSAs shown in Box 5.1.1 included important public health targets under objective 1, such as improved reduced mortality rates from heart disease and cancer; reductions of health inequalities; and reduced rates of smoking, childhood obesity and teenage pregnancy. Converting these high-level public health objectives into meaningful local targets through the medium of the performance ratings system proved far less straightforward than in the waiting time domain. Public health has not received anything like the sustained managerial attention enjoyed by the health service delivery targets (Marks & Hunter 2005). This raises concerns that local managers concentrated on targeted and readily managed aspects of health care (most notably objective 3 – waiting times) at the expense of less controllable and less immediate concerns, such as public health (objective 1).

Whilst retaining the principle of rating performance on a simple composite measure, it is noteworthy that in 2006 the Healthcare

Commission implemented a major change to the assessment regime that pays more attention to a broader spectrum of performance, most notably clinical quality (Healthcare Commission 2005). This places greater emphasis on self reporting and reports clinical performance and financial performance separately.

Discussion

PSAs and, in particular, the associated targets have become a central element of political discourse in England. Without question they have succeeded in shaping the priorities and delivery of public services in general, and health services in particular, although it remains a matter of fierce debate whether that influence is for the good. On the one side are those who claim that their focus on outcomes and setting of firm measurable targets have helped to modernize those services. On the other are those who claim that their simplistic view of priorities has undermined the traditional public service ethos and rendered those services dysfunctional.

In the health domain PSA targets have certainly delivered noteworthy successes, such as the reduction in NHS waiting times. However, alongside the manifest intended improvements in many of the measured PSA targets there are widespread reports of adverse side effects in other, often unmeasured, aspects of public services (Bevan & Hood 2006a). Many of these reports are anecdotal and may be apocryphal, but some have been credibly documented by the House of Commons Public Administration Select Committee (2003) and Bevan and Hood (2006b). These include neglect of unmeasured aspects of performance (e.g. sacrificing clinical priorities in the pursuit of reduced waiting times); distorted behaviour (e.g. refusing to admit patients to accident departments until a four-hour waiting time target was achievable); and fraud (manipulation of waiting lists).

Unintended and adverse responses such as these were readily predictable from the Soviet literature (Nove 1980). They offer a powerful caution against relying solely on a targets regime to secure improvement and indicate the need for countervailing instruments (Smith 1995). These might include: strong national data audit and surveillance capacity; system of professional inspection that monitors and reports on unintended consequences; careful scrutiny of performance beyond targets by organizational boards of governors; some sort of

democratic 'voice' in the control of local public service organizations; and empowerment of service users through improved information and systems of redress.

The Social Market Foundation (2005) summarized the criticisms of the targets regime under five headings:

1. there are too many targets
2. they are too rigid and undermine the morale of staff
3. they have perverse and unintended consequences
4. not always clear who is responsible for meeting the target
5. data are often not credible.

Over its ten-year lifetime, the PSA infrastructure has been adapted as difficulties have arisen and remedial measures put in place. Drawing on the experience of other countries (where available) this section discusses some of the most important questions that have arisen in the development of the English system under eight headings: (i) Who should choose the targets? (ii) What targets should be chosen? (iii) When should outcomes be used as a basis for targets? (iv) How should targets be measured and set? (v) How should cross-departmental targets be handled? (vi) How should attainment be scrutinized? (vii) How should departmental objectives be transmitted to local organizations?

Who should choose the targets?

In principle, it seems perfectly reasonable (and indeed honourable) for a legitimately elected government to set out its objectives and targets in an explicit fashion. Targets serve many purposes, one of which is to enhance political accountability. Indeed, lack of an adequate accountability framework may lead to failure to achieve the objectives of target setting (see Box 5.1.3). The PSAs enable parliament and the electorate to hold the government to account for both its choice of priorities and its performance against the targets. Indeed, it is a sign of the success of the process that much of the public debate surrounding targets referred less to the principle of setting targets and more to the details of what they should be.

However, disagreement remains about the processes by which priorities are chosen and targets are set. For example, many argue that the government's excessive emphasis on waiting times in NHS targets has posed a threat to clinical quality by ignoring the prime objective of

Box 5.1.3 Lack of accountability in Hungarian target setting

In Hungary, the lack of an accountability framework was identified as one of the reasons why target setting failed to achieve its objectives. Political will served as the sole determinant of whether or not a health policy would be target-based. Ten years after the development of the first target-based health policies, there is still no legal pressure to develop the policies further (Vokó & Ádány 2008). The following have been recognized as contributory factors in the failure to establish an accountability framework.

1. An overall feeling of lack of ownership resulted from the realization that the Hungarian health monitoring system was capable of providing information only at the national level and thus could not take account of huge social and geographical inequalities.
2. Policy-makers and those involved from outside the health sector were rarely involved in the development of the targets. An interministerial committee was set up to coordinate the targets and to try to bridge the gap between the various sectors but its work was hindered by the very limited financial resources allocated to targets in Hungary.
3. Slow acceptance of the new public health approach in Hungary reflected a lack of awareness among health professionals. Health is not a priority issue for other sectors and so they were reluctant to incorporate health considerations into their own policies.

As a result, Hungarian targets lack regulation, ownership, consensus and financing and have failed to induce behaviour change.

Source: Vokó & Ádány 2008

health care – to improve health. Such outcomes have led some to argue that the professionals who deliver the public services should have a greater say in the nature of the targets. There is an element of good sense in this principle, especially in health services where outcomes rely very heavily on the engagement and commitment of front-line professionals. Yet it is also the case that the priorities and working practices of those professionals may impede progress towards better performance. To some extent, the PSA process seeks to challenge tra-

ditional ways of delivering services and therefore at times will come into conflict with the professions.

Some argue that parliament should have a greater say in target setting. Parliament already plays a crucial role by scrutinizing the choice of priorities and the attainment of targets and it is difficult to see how the legislature's involvement in choosing targets would enhance the PSA process. No government will pursue objectives with total commitment when it does not fully control their nature. Of course, this also applies to the devolved organizations charged with delivering services and gives rise to some of the problems of morale and alienation discussed below.

It is frequently suggested that service users should have more say in setting PSA targets and of course there is much to commend wide consultation with user groups when identifying priorities for improvement. However, the setting of objectives involves considerations beyond immediate users of a particular service, such as the taxpayer perspective; the interests of future users; and the interests of users of other services. The user perspective is important but cannot be the sole influence on priority setting, which in any case involves judgements about the relative importance of different user groups.

Consensus and ownership have nevertheless usually been seen as vital to elicit acceptance of country-based targets. In Catalonia, health councils were created at central and provincial levels to encourage citizens' groups to take an active part in target setting. In Flanders, Belgium, local health networks (LHN) were established to encourage the exchange of information between local organizations and create possibilities for collaboration by offering a focal point for preventive actions. The organizations were encouraged to undertake collaborations with local government and other sectors to achieve the health targets (Van den Broucke 2008). France saw the establishment of national and regional health conferences which allowed stakeholders the opportunity to debate existing health problems and foster partnerships. It is clear that targets without consensus and ownership will have difficulty achieving success.

In isolation, neither consensus and ownership nor legislation can guarantee results. Implementation of an accountability framework demands vertical and horizontal coordination, which can be difficult. In Flanders, the five health targets were repeatedly reaffirmed between 1998 and 2003 when a decree was passed to outline the procedures

for formulating new targets and updating existing ones. This decree helped to streamline the process of target setting and provided a legal basis for synchronizing the activities of the different players involved. Ten years after the first targets were introduced in Flanders, it is clear that health targets have become a well-established mechanism to support prevention policies. They may not have produced the anticipated results in terms of health gain or changes in health-related behaviour but they have spurred changes in the policy environment that may assist in achieving targets in the future.

Hence, any prudent government seeking to implement a PSA type process would be well-advised to consult many relevant stakeholders to reach consensus on the choice of objectives and the nature of the targets. However, uncritical accommodation of every interest group would render the target process meaningless, for example by leading to an unwieldy proliferation of priorities. One of a government's prime roles is to balance conflicting claims on public resources and targets should be an explicit and succinct statement of the government's decisions.

What targets should be chosen?

Multiple objectives are a characteristic of health services. Indeed, it can be argued that the existence of multiple objectives is one of the defining characteristics of public services such as health care and one of the reasons why they cannot (at least in their entirety) be delivered by competitive markets.

However, one intention of any targets regime is to focus on a limited number of objectives. The initial 1998 suite of English PSAs failed to recognize that this requires tough political choices and therefore failed to have a detectable impact in many domains. This mistake was not confined to England but visible in many other target programmes developed around that time such as the 1998 programmes in Italy (100 targets) and in Andalucia (84 targets) (Busse & Wismar 2002). Subsequent English spending reviews addressed this issue by focusing on a greatly reduced number of targets. Nevertheless, it is important to note that some of the numerical reduction was deceptive – the 2004 example given above indicates how some targets became multidimensional, for example seeking to address both overall health improvement and reductions in inequalities in health. Also as noted above a

number of previous targets were converted into standards, indicating a level of attainment secured in previous periods. To many, these retained the appearance of targets albeit in a different guise.

Having identified a priority, it is noteworthy that the English government sought to include an associated objective into the targets regime, even when attainment is hard to measure (e.g. patient experience target in Box 5.1.1.) Without question quantification is a good principle to pursue as it generally allows the government to set concrete targets for departments. However, it runs the risk of distracting managerial attention from important qualitative aspects of performance and suggests that reports of progress towards quantified targets should be accompanied by a narrative that describes success and failure in more qualitative terms.

The move towards specifying standards indicates that targets should focus on domains where manifest change is required, as the Social Market Foundation suggests. If a domain is not included in the targets regime, this does not necessarily indicate that it is unimportant. Rather, it may suggest that it is not a priority for urgent change and should instead be considered a standard. The key focus of targets should be where change is required and maintenance of standards in other domains should be secured through other instruments, such as routine regulation, inspection or market mechanisms.

When should outcomes be used as a basis for targets?

From the outset, the architects of the English targets system recognized that the outcomes of public services usually matter to most service users and the broader public. In principle, the outcomes focus should enable health service organizations to look beyond traditional ways of delivering their services and traditional organizational boundaries.

However, the focus on outcomes can give rise to difficulties. For example, some outcomes (e.g. many aspects of health system responsiveness) are intrinsically difficult to measure. Even if they can be measured, some outcomes (e.g. reduced mortality from smoking) can take years to materialize – beyond the lifetime of most governments. Furthermore, some outcomes (e.g. most conventional mortality rates) are particularly vulnerable to influences beyond the control of the health ministry. Each of these difficulties offers the ministry an excuse for apparent failure and can undermine the targets process.

On the other hand, it is clear that the use of process measures can distort behaviour and lead to unintended outcomes. For example, the attempt to guarantee access to a primary care professional within twenty-four hours led to widespread reports of primary care practices refusing to allow patients to arrange appointments more than twenty-four hours in advance, even when that was their preference. Patients could secure access to appointments only by telephoning on the day they required a consultation, often leading to uncertainty and inconvenient timing of appointments. The real objective (securing quicker and more convenient access to a doctor) was subverted by the use of an incomplete and poorly articulated target. Thus, if the chosen output target is pursued without regard to the eventual outcomes, additional assurance will be needed to ensure that the desired outcomes have indeed been secured. In this example, if the real objective was to increase patient satisfaction, it would have been preferable to use a direct measure of patient satisfaction (rather than a highly imperfect proxy measure) as the basis for the target.

In short, outcome measures address what matters to the service user and the citizen and are less vulnerable to distortion. It therefore seems incontestable that outcomes should inform all targets. However, there will be occasions when a carefully chosen output or process measure – which evidence shows to be clearly linked to the eventual outcome – may form a more effective basis for a target.

How should targets be measured and set?

A central feature of the English targets debate has been how (once objectives have been identified) the associated targets should be set, in terms of the required measurement instrument and level of attainment. The use of SMART targets was advocated in the United Kingdom (HM Treasury et al. 2001) as in other countries and the Treasury has sought to pursue these principles when setting PSA targets.

The Royal Statistical Society (Bird et al. 2005) put forward a more comprehensive set of desirable general principles for setting targets.

- Indicators should be directly relevant to the primary objective, or be an obviously adequate proxy measure.
- Definitions need to be precise but practicable.

- Survey-based indicators, such as those of user satisfaction, should use a shared methodology and common questions between institutions.
- Indicators and definitions should be consistent over time.
- Indicators and definitions should obviate, rather than create, perverse behaviours.
- Indicators should be straightforward to interpret, avoiding ambiguity about whether the performance being monitored has improved or deteriorated.
- Indicators that are not collected for the whole population should have sufficient coverage to ensure against misleading results, that is: potential bias compared to measuring the target population should be small.
- Technical properties of the indicator should be adequate.
- Indicators should have the statistical potential to exhibit or identify change within the intended timescale.
- Indicators should be produced with appropriate frequency, disaggregation and adjustment for context.
- Indicators should conform to international standards if these exist.
- Indicators should not impose an undue burden – in terms of cost, personnel or intrusion – on those providing the information.
- Measurement costs should be commensurate with the likely information gain.

The National Audit Office (2005 and 2006) scrutinized the data systems used to monitor and report progress against all 2002 PSA targets and found varying levels of success – only 30% were deemed strictly fit for purpose. The Statistics Commission (2006) scrutinized all 2004 targets in detail to assess whether the statistical evidence to support PSA targets was adequate for the purpose of achieving government policy objectives. It noted numerous problems with poor specification; undue complexity; and availability, transparency, independence and timeliness of the data.

A number of approaches exist to overcome some of these weaknesses. For example, the Royal Statistical Society advocates a multistage measurement 'protocol' for each target that would explicitly explain all stages of the measurement process, from choice of indicator to publication of results (Bird 2005). It also recommends publica-

tion of levels of uncertainty alongside all attainment measures. In the same vein, the Statistics Commission (2006) advocates publication of interim attainment measures for longer-term targets.

The specification of explicit levels of attainment is a particular feature of targets regimes. However, this important element of the process is usually applied with inconsistent rigour. Some targets might be little more than unattainable aspirations whilst others can be secured with little effort on the part of ministries. Furthermore, there are conflicting pressures within any targets regime. To be effective managerial instruments, targets should be stretching but attainable, suggesting (say) a one in three risk of failure. However, few governments would want to be confronted with such a high proportion of failures. From an accountability perspective, they would wish to feel that there was a good chance of attaining all targets.

This was seen in the Netherlands during the early 1990s when the Secretary of State for Health avoided using quantitative health targets because of the political accountability that they would create (van Herten & Gunning-Schepers 2000). Similarly, Russia has experienced politically driven target setting in which the targets set were neither especially relevant nor necessary. Health was seldom a priority on the policy agenda in the USSR or subsequently in the Russian Federation and generally those targets that were set were broadly defined, infrastructure-oriented and almost never outcome-oriented. In many cases, achievement of the targets required no change in policy (Danishevski 2008). It is difficult to see how this tension can be resolved satisfac-torily as it requires a political process mature enough to recognize that some failure is inevitable and not necessarily adverse if progress is also being secured.

A note of caution is helpful in this context. A target that is not achieved is easily dismissed as 'too ambitious' (as in the Netherlands); a target that is achieved is sometimes dismissed as 'would have been reached anyway' (e.g. coronary heart disease death rate target in England). These deserve closer examination. The first statement requires a thorough knowledge of the potential effect sizes (efficacy) of various intervention strategies (and possible combinations of interventions); the second assumes that longitudinal trends remain constant over time. This is not the case since external factors also exercise large influences.

Life expectancy in Central and Eastern Europe provides a good example of this point. If, in 1990, Russia had passed a target to keep life expectancy constant until 2000, it would have been accused of set-

ting a target that would be reached anyway. In reality, if this target had resulted in halving the actual decline it would have been a success even though most evaluation strategies would label it a failure. The same holds true in reverse. If a target that 'experts' have judged to be neither overambitious nor trivial has been reached successfully, it is rather difficult to attribute this to the strategy itself. This argues for an independent assessment of the attribution of success or failure. However, it is usually not possible to differentiate with any confidence how the different elements have contributed to the measured outcome and we shall probably never be able to control for all factors contributing to good or ill health (Busse 1999).

How should cross-departmental targets be handled?

The many determinants of health involve actions by organizations in many different sectors and effective coordination among responsible actors has emerged as a key issue in securing system improvements. In particular, a focus on health outcomes sometimes gives rise to strategies that are not obviously attached to a particular ministry, leading to the need to specify joint targets that transcend departmental boundaries. These are particularly important in the public health domain and have produced difficulties in the English PSA process. A joint report by the National Audit Office and the Audit Commission (2006) examines complex cross-departmental targets, including efforts to halt the rise in child obesity. They find no ready solutions, but advocate much stronger collaboration between national and local government and stronger engagement with non-governmental organizations.

In short, cross-sectoral targets give rise to problems of coordination, persuasion and engagement that must be addressed if they are to be successful. Effective coordination depends on the structures already in place, particularly the system of governance and the forums within which key actors can meet. This may be easier where responsibility for health lies within local or regional government, as in Scandinavia, but it is possible to convene relevant actors from many different sectors in other ways.

The Social Market Foundation (2005) recognizes that some targets cannot be broken down into individual components and therefore require joint effort by two or more ministries. However, it recommends that there should always be a 'lead' ministry that takes responsibility for meeting the target. It is noteworthy that the 2007 Comprehensive

Spending Review placed special emphasis on cross-departmental collaboration, with a view to seeking innovative solutions for the challenges posed by joint targets.

Other countries have faced a different challenge with intersectoral targets. Having stressed the need to involve the many sectors whose actions contribute to health, often they have not included the health-care sector itself. This has made health targets an issue for actors only at the sideline, thereby often diluting their potential impact (Busse & Wismar 2002).

How should attainment be scrutinized?

A persistent theme in any discussion of targets is how to scrutinize, understand and report on progress. Given that this mechanism has played such a central part in the recent development of English public services, there has been surprisingly little attention to public reporting and scrutiny of attainment against targets.

One exception was the House of Commons Public Administration Select Committee (2003) report which sought to identify attainment of 249 measurable targets from 1998. These results were not readily available but research revealed that 67.1% had been met; 7.6% partially met; 10.0% not met; and 14.9% had inadequate data on achievement. In their original form, performance reports were found in a variety of formats and with varying levels of clarity in the annual reports of individual ministries. The Treasury web site merely offered links to these reports. The Committee recommended that progress towards targets and eventual attainment should be reported consistently and regularly on a single, authoritative web site.

Within many parliamentary systems, the parliament appoints scrutiny committees for most ministries. These would be the natural focus for holding a government to account through routine reporting of progress towards targets. However, systematic parliamentary scrutiny has not yet become routine in England and the Health Select Committee has referred to PSA targets only periodically. Thus, scrutiny has been piecemeal – e.g. in the form of occasional reports from pressure groups, the media and regulators.

Within any targets regime it is particularly important to ensure independent audit of the reliability of the data used to assess attainment. Within government, few have an interest in challenging information

that reports apparent performance improvements and attainment of targets. In England this has led to considerable popular scepticism about the veracity of information that the government provides on its own performance. The National Audit Office examines the processes for data collection but is not in a position to assure the accuracy of all data. It is noteworthy that the British Government has made the Office for National Statistics more independent of government by creating the UK Statistics Authority, accountable directly to parliament. An important objective of this initiative is to dispel the perception that reports of government performance may be unreliable.

Within government there has been far greater attention to scrutiny of progress towards English targets. Service delivery agreements with departments were the initial instruments for assuring the implementation of PSA targets. When these proved unwieldy and ineffective they were replaced by the Prime Minister's Delivery Unit, a very important element of the more mature PSA system. This indicates a perception within government that continuous monitoring; strong and timely intervention powers; and continued political attention at the highest level have made essential contributions to the longevity and sustained high profile of the system.

How should departmental objectives be transmitted to local organizations?

Attainment of national ministerial targets usually relies on securing satisfactory improvement in local organizations charged with the delivery of health services. Therefore, much depends on how ministerial targets are transmitted to local services. For example, it would be clearly inappropriate to set the same mortality targets for every locality, regardless of existing levels of attainment and the difficulty of local circumstances. Such approaches lead to manifest problems. Organizations that are already performing well have no incentive to improve; those with disadvantaged populations may stand no chance of success and become alienated. Indeed, if such regimes are sustained, existing problems may be exacerbated as it becomes difficult to recruit key managers and professionals in disadvantaged areas. As a result, many ministries have introduced more subtle target regimes for local organizations and sought to encourage all organizations to improve in the chosen measures, whatever their baseline.

The tension between national objectives and local discretion has become an important unresolved issue within the English regime. In particular, the 'must do' nature of local health targets has put especially severe pressure on some local organizations, precluding any serious consideration of separate local priorities. The prevailing lack of flexibility was highlighted in a report by the Audit Commission (2003) that criticized the neglect of local government discretion in earlier PSA targets. The Treasury responded by setting up a review of devolved decision-making to examine how national priorities and local flexibility can be accommodated within the targets system. It is moving towards the publication of local performance data as an alternative to national targets (HM Treasury 2004b). The aim is to allow local people (rather than national government) to hold local services to account for their chosen priorities and performance. However, whilst a policy of devolution clearly has relevance to health systems delivered through local governments, it is not clear how local accountability can be secured in health systems that do not have a local democratic decision-making mechanism.

This problem is not confined to England. All countries need to develop a sense of ownership and accountability amongst those required to implement health targets. Unfortunately, this is often not the case. As a previous review has noted, target programmes are disseminated in a top-down manner with little effort to ensure the involvement of key actors at the grass-roots level (Wismar & Busse 2002).

Conclusions

The use of targets is becoming widespread in health systems and therefore is clearly perceived to be an important mechanism for securing health system improvement and accountability. In particular, the English health system's experience with targets has developed very rapidly over a period of fifteen years. The first tentative steps in the domain of public health were largely ineffective and initial ambitions were modest when attention switched to health service delivery. However, the introduction of a targets 'culture' throughout English public services rapidly increased the prominence and impact of targets in the NHS, most notably in the form of performance ratings of NHS organizations.

The government had a number of objectives when it introduced the PSA system in 1998 (House of Commons Public Administration Select Committee 2003):

- to offer a clear statement of what it is trying to achieve
- to give a clear sense of direction and ambition
- to introduce a focus on delivering results
- to form a basis for deciding what is and what is not working
- to improve accountability.

It is difficult to argue with the claim that, at least in parts of the health domain, the PSA system has been successful in these respects. Smith (2008) suggests a number of reasons for the increasing influence of targets. First, their range and specificity has increased markedly – moving from long-term general objectives towards very precise short-term targets. Second, the specification has moved progressively from the national to the organizational level. This local interpretation of national targets is likely to have much more resonance with local decision-makers. Third, some attempts have been made to engage professionals with the design and implementation of the targets regime. This runs the risk of capture by professional interests but also increases the chance that professionals will take notice of the targets. Fourth, organizations have been given increased capacity to respond to challenging targets, in the form of extra finance, information and managerial expertise. Finally, very concrete incentives have been attached to the targets.

It is noteworthy that the English target initiatives have in effect combined a multiplicity of targets into a single indicator of performance at the local level (the performance ratings). As discussed in Chapter 3.4, if the method of aggregating individual indicators is in line with national objectives then these composite measures of success can play a particularly important role in capturing the attention of local decision-makers and allowing local organizations to choose the areas of endeavour that they wish to concentrate on. The alternative – requiring improvement in every domain – diminishes such local autonomy and may be less effective.

The use of targets remains a work in progress that has introduced numerous challenges and anomalies, as documented in this chapter. As experience unfolds, it is becoming clear that a targets regime must be augmented by a number of other mechanisms. In a series of depart-

mental Capability Reviews by the Cabinet Office (2006) in the United Kingdom it was noted that '… whilst progress against PSAs and other top targets is necessary and welcome, it is not sufficient for delivering high-quality performance across the whole system.'

Some of the more important institutional requirements for the implementation of regimes such as the English PSA system are listed below.

- Sustained political commitment to the targets system, at the very highest level.
- A nimble central government organization (Prime Minister's Delivery Unit) responsible for timely monitoring, reporting and (where necessary) intervention.
- Continued monitoring and regulation in domains not directly covered by targets.
- High-quality performance management skills within the ministry.
- Carefully crafted mechanisms for transmitting national targets to the local level.
- Strong collaborative arrangements, where necessary, for domains that cross traditional ministerial boundaries.
- Careful integration of central and local priorities.
- Engagement as appropriate with relevant stakeholders, including user groups, professional organizations and the voluntary sector.

A number of commentators have offered suggestions on the architecture of the targets regime. For example, the Social Market Foundation (2005) raises several issues.

- Targets should be set only when change is required or for aspects of public services which are exceptionally important.
- There should be a fairly small number of targets in place at any one time.
- Whilst an outcome orientation is desirable, process and input targets may sometimes be appropriate, especially if the organization in question has limited influence over the outcome.
- Targets add most value where other mechanisms such as user choice and the threat of exit, or the contestability of providers, are not in place.
- Proportionate sanctions and incentives are important. An organization that misses a stretching target by a narrow margin should not be sanctioned for failure, but rather rewarded for its progress.

- Targets should be fully integrated into ministerial performance management, audit and inspection regimes.
- Joint targets that need to be delivered by more than one department should always have a lead ministry that takes responsibility for meeting the target.
- Greater use could be made of targets relating to public satisfaction.

In addition, the Royal Statistical Society and the Statistics Commission have given detailed guidance on technical aspects of performance measurement (Bird et al. 2005; Statistics Commission 2006). The work of the National Audit Office emphasizes the need to improve data quality and there is clear evidence that genuinely independent scrutiny and audit of the data has become a central requirement of any targets regime.

Notwithstanding a cautiously positive commentary on recent English experience with targets in health, Smith (2008) has noted some serious risks drawn from the English experience, including those listed here.

- Targets are selective and untargeted aspects of the health system may suffer from neglect.
- Unless incentives are designed carefully, managers and practitioners are likely to concentrate on short-term targets directly within their control at the expense of targets addressing longer-term or less controllable objectives.
- The targets system is very complex, requiring capacity to implement and giving rise to the scope for capture by professional interests.
- Excessively aggressive targets may undermine the reliability of the data on which they depend.
- Excessively aggressive targets may induce gaming or other undesirable labour market responses, as clinicians seek to create favourable environments for achieving those targets.
- The targets regime may replace altruistic professional motivation with a narrow mercenary viewpoint.

A full evaluation of the costs and benefits of any English targets system is likely to be intrinsically difficult and is still awaited. However, most of the risks can be mitigated to some extent by careful monitoring and the introduction of countervailing instruments where necessary. Targets have secured a real change in the behaviour of the English

health system, probably to a much greater extent than any previous policy instruments. The challenge for any health system that relies on targets is to monitor carefully; to nurture the benefits of targets; and to neutralize their harmful side-effects.

References

Audit Commission (2003). *Targets in the public sector*. London: The Stationery Office (http://www.audit-commission.gov.uk/reports/NATIONAL-REPORT.asp?CategoryID=&ProdID=B02E376A-01D5-485b-A866-3C7117DC435A).

Barker, R. Pearce, M. Irving, M (2004). 'Star wars, NHS style.' *British Medical Journal*, 329(7457):107–109.

Bevan, G. Hood, C (2006). 'What's measured is what matters: targets and gaming in the English public health care system.' *Public Administration,* 84(3): 517–538.

Bevan, G. Hood, C (2006a). 'Have targets improved performance in the English NHS?' *British Medical Journal,* 332(7538): 419–422.

Bird, SM. Cox, D. Farewell, VT. Goldstein, H. Holt, T. Smith, P (2005). 'Performance indicators: good, bad and ugly,' *Journal of the Royal Statistical Society, Series A*, 168(1): 1–25 (http://www.rss.org.uk/PDF/PerformanceMonitoring.pdf).

Busse, R (1999). 'Evaluation and outcomes of health targets.' *Eurohealth*, 5(3):12–13.

Busse, R. Wismar, M (2002). 'Health target programmes and health care services - any link? A conceptual and comparative study (part 1).' *Health Policy*, 59(3): 209–221.

Cabinet Office (2006). *Capability reviews: the findings of the first four reviews*. London: Cabinet Office (http://www.civilservice.gov.uk/reform/capability_reviews/reports.asp).

Danishevski, K (2008). The Russian Federation: difficult history of target setting. In: Wismar, M. McKee, M. Ernst, K. Srivastava, D. Busse, R (eds.). *Health targets in Europe: learning from experience*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

Department of Health (1992). *The Health of the Nation*. London: HMSO.

Department of Health (1998). *Health of the Nation: a policy assessed*. London.

Department of Health (2001). *NHS performance ratings Acute Trusts 2000/01*. (http://www.doh.gov.uk/performanceratings/).

Hauck, K. Street, A (2007). 'Do targets matter? A comparison of English and Welsh national health priorities. *Health Economics*, 16(3): 275–290.

Healthcare Commission (2005). *Assessment for improvement. The annual health check. Measuring what matters*. London: The Healthcare Commission (http://annualhealthcheckratings.healthcarecommission.org.uk/annualhealthcheckratings.cfm).

HM Treasury (2004). *Stability, security and opportunity for all: investing for Britain's long-term future. New public spending plans 2005-2008*. London: The Stationery Office (http://www.hm-treasury.gov.uk/spending_review/spend_sr04/spend_sr04_index.cfm).

HM Treasury (2004a). *Devolving decision making: 1- delivering better public services: refining targets and performance management*. London: The Stationery Office (http://www.hm-treasury.gov.uk/media//53886/devolving_decision1_409.pdf).

HM Treasury, Cabinet Office, National Audit Office, Audit Commission, Office for National Statistics (2001). *Choosing the right fabric: a framework for performance information*. London: HM Treasury (http://www.hm-treasury.gov.uk/media/EDE/5E/229.pdf).

House of Commons Public Administration Select Committee (2003). *On target? Government by measurement. Fifth Report of Session 2002–03*. London: House of Commons (http://www.publications.parliament.uk/pa/cm200203/cmselect/cmpubadm/62/6202.htm).

Hunter, D (2002). England. In: Marinker, M (ed.). *Health targets in Europe*. London: BMJ Books.

Mannion, R. Davies, H. Marshall, M (2005). 'Impact of star performance ratings in English acute hospital trusts.' *Journal of Health Services Research and Policy*, 10(1): 18–24.

Marks, L. Hunter, D (2005). 'Moving upstream or muddying the waters? Incentives for managing for health.' *Public Health*, 119(11): 974–980.

National Audit Office (2005). *Public service agreements: managing data quality – compendium report*. London: The Stationery Office (http://www.nao.org.uk/pn/04-05/0405476.htm).

National Audit Office and Audit Commission (2006). *Delivering efficiently: strengthening the links in public service delivery chains*. London: The Stationery Office (http://www.nao.org.uk/pn/05-06/0506940.htm).

Nove, A (1980). *The Soviet economic system, second edition*. London: Allen and Unwin.

Propper, C. Sutton, M. Whitnall, C. Windmeijer, F (2008). 'Did "targets and terror" reduce waiting times in England for hospital care?' *B.E. Journal of Economic Analysis & Policy,* 8(2): 1863.

Ritsatakis, A. Barnes, R. Dekker, E. Harrington, P. Kokko, S. Makara, P (2000). *Exploring health policy development in Europe*. Copenhagen: WHO Regional Office for Europe.

Smith, P (1995). 'On the unintended consequences of publishing performance data in the public sector.' *International Journal of Public Administration*, 18(2/3): 277–310.

Smith, PC (2008). England: intended and unintended effects. In: Wismar, M. McKee, M. Ernst, K. Srivastava, D. Busse, R (eds.). *Health targets in Europe: learning from experience*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

Social Market Foundation (2005). *To the point: a blueprint for good targets*. *Report of the Commission on Targets in the Public Services*. London: Social Market Foundation.

Statistics Commission (2006). *PSA targets: the devil in the detail*. London: Statistics Commission (http://www.statscom.org.uk/media_pdfs/reports/Final%20PSA%20Targets%20Report.pdf).

Van den Broucke, S (2008). Flanders: health targets as a catalyst for action. In: Wismar, M. McKee, M. Ernst, K. Srivastava, D. Busse, R (eds.). *Health targets in Europe: learning from experience*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

Van de Water, HPA. van Herten, LM (1998). *Health policies on target? Review of health target and priority setting in 18 European countries*. Leiden: TNO.

van Herten, LM. Gunning-Schepers, LJ (2000). 'Targets as a tool in health policy. Part I: lessons learned.' *Health Policy*, 53(1): 1–11.

Vokó, Z. Ádány, R (2008). Hungary: targets driving improved health intelligence. In: Wismar, M. McKee, M. Ernst, K. Srivastava, D. Busse, R (eds.). *Health targets in Europe: learning from experience*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

WHO Regional Office for Europe (2005). *The Health for All policy framework for the WHO European Region: 2005 update*. Regional Committee for Europe. Fifty-fifth session. Copenhagen: WHO Regional Office for Europe (http://www.euro.who.int/Document/RC55/edoc08.pdf).

Wismar, M. Busse, R (2002). 'Outcome-related health targets – political strategies for better health outcomes. A conceptual and comparative study (part 2).' *Health Policy*, 59(3): 223–241.

Wismar, M. McKee, M. Ernst, K. Srivastava, D. Busse, R (eds.) (2008). *Health targets in Europe: learning from experience*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

5.2 *Public performance reporting on quality information*

PAUL G. SHEKELLE

Introduction

The public reporting of information about the quality of health care delivered by identified providers has become increasingly popular in developed countries. In part this is due to a general trend towards increasing the transparency of the performance of a variety of services (e.g. test scores in schools). Within health care this is also promoted as a mechanism to help improve the quality of care. Berwick et al's (2003) framework for quality improvement shows that public reporting can improve quality via two pathways. In the first (selection pathway), consumers (patients) select providers of better quality. In the second (change pathway), performance data help providers to identify areas of underperformance and public release of the information acts as a stimulus for improvement (Fig. 5.2.1).

Colleagues and I recently completed a systematic review of the published evidence regarding the public release of performance data to improve quality, identifying forty-five articles (Fung et al. 2008). This

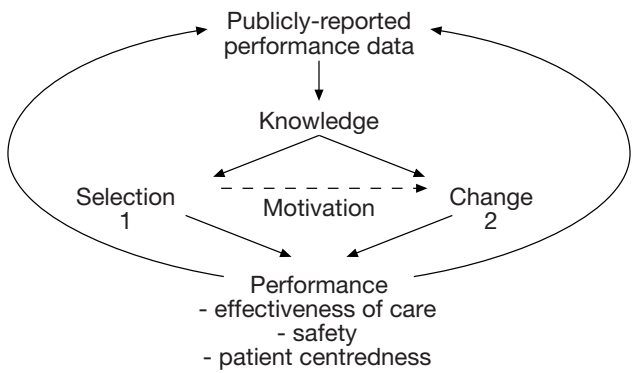


Fig. 5.2.1 Two pathways for improving performance through release of publicly-reported performance data

Source: Berwick et al. 2003

updated the earlier review on the same topic (Marshall et al. 2000). In this chapter I discuss the evidence from these reviews in the context of key questions and conclusions for the WHO conference.

Public reporting: effect on selection pathway

There is evidence that public reporting has little effect on the selection pathway. In our review, we identified twenty-one studies that assessed the effect of public reporting on the selection of health plans, hospitals or providers. Studies were mostly observational in design, being time series analyses of market share during the period of the introduction of public reporting. Experimental studies of consumers' response to hypothetical quality ratings revealed some willingness to trade access restrictions for higher quality (Harris 2002; Spranca et al. 2000). However, two randomized trials of Medicare beneficiaries' use of data from the Consumer Assessment of Healthcare Providers and Systems (CAHPS) in the United States showed that this public reporting had no overall effect on the selection of health plans (Farley et al. 2002 & 2002a). We know of one other randomized trial of the effect of the release of actual quality information on health plan selection – this has not yet been published.

For hospitals, nine studies of four different American public reporting systems showed no or (at most) modest short-term effects on market share (Baker et al. 2003; Chassin 2002; Hannan et al. 1994a; Hibbard et al. 2005; Jha & Epstein 2006; Mennemeyer et al. 1997; Mukamel & Mushlin 1998; Vladeck et al. 1988). For example, two analyses of one of the earliest public reporting systems – the Health Care Financing Administration (now CMS) release of hospital mortality rates – reported statistically significant but small changes in utilization (Mennemeyer et al. 1997) or no statistically significant changes in bed occupancy rates (Vladeck et al. 1988).

Among studies that assessed the effect on market share of the New York State CSRS, three out of four concluded that effects (if any) were minimal (Chassin 2002; Hannan et al. 1994a; Jha & Epstein, 2006; Mukamel & Mushlin 1998). We found seven studies regarding individual providers. Public reporting of performance data was associated with ceasing practice for low volume cardiac surgeons in the New York State CSRS, but other effects were small or inconsistent (Hannan et al. 1994a & 1995; Jha & Epstein 2006; Mukamel & Mishlin 1998; Mukamel et al. 2000, 2002 & 2004–2005).

Public reporting: effect on quality improvement activities (change pathway)

By hospitals

There is good evidence that public reporting stimulates quality improvement activities by hospitals (change pathway). We identified eleven studies, almost all of which found that the public release of performance data stimulated activities at the hospital level. For example, a controlled trial by Hibbard and colleagues showed that the quantity of quality improvement activities was greater in hospitals subject to public reporting than in those receiving confidential reporting of the same quality information (Hibbard et al. 2003 & 2005). Similarly, Tu and Cameron (2003) found that more than half of the hospitals responded to a Canadian hospital-specific report on acute myocardial infarction by implementing quality improvement activities. Chassin (2002) conducted a series of interviews and case studies that documented the steps taken to try to improve cardiac surgery programmes within New York hospitals. Other studies reported similar findings – hospitals acted in response to public reporting of performance data (Bentley & Nash 1998; Dziuban et al. 1994; Longo et al. 1997; Mannion et al. 2005; Rosenthal et al. 1998). For example, Rosenthal et al. (1998) assessed hospitals participating in the Cleveland Health Quality Choice programme. Examining one academic and three community hospitals, they found increases in quality improvement activities such as interdisciplinary process improvement teams; review of processes of care; and development of practice guidelines. Only two studies reported that public reporting had little effect on hospital activity, both concerned the same system – the California Hospital Outcomes Project (Luce et al. 1996; Rainwater et al. 1998).

By health plans or individual providers

We identified no studies that assessed the effect of the public reporting of performance information on quality improvement activities by health plans of individual providers. However, the changes observed in hospitals are expected to carry over to health plans and individual providers and there are nonsystematic data about the changes instituted by health plans in order to improve performance on public

quality measures. For example, a recent commentary on performance measurement reported that an American insurance company and health plan (Aetna) developed a plan to respond to the HEDIS requirement to use the administration of beta blockers following myocardial infarction as a performance measure. The use of beta blockers was integrated within the 'scripts' used by their case managers following Aetna members who had suffered a myocardial infarction. The company also started to send information about beta blockers to patients and their physicians (Lee 2007).

It is likely that the lack of published studies documenting the effect of public reporting on quality improvement activities by health plans or individual providers is due not to any lack of effect but rather because this is happening outside the usual sphere in which academic physicians work, research and publish.

Public reporting: effect on clinical outcomes

There is scant direct evidence that public reporting improves clinical outcomes. Without doubt, the greatest number of published studies about the effects of the public release of performance data concern mortality associated with cardiac surgery, specifically the New York State CSRS. Eight studies assessed the effect of public reporting on hospital clinical outcomes focused on the CSRS (Dranove et al. 2003; Dziuban et al. 1994; Ghali et al. 1997; Hannan et al. 1994 & 1994a; Moscucci et al. 2005; Omoigui et al. 1996; Peterson et al. 1998). All are in agreement that there has been a marked decline in mortality during the time that the CSRS has been in place. The issue is whether this decline is greater than in other areas of the United States that have no public reporting (i.e. is a secular trend unassociated with the CSRS) or whether the decline is due to New York cardiac surgeons' avoidance of high-risk patients and/or outmigration of such cases to other states. Suffice to say that this issue has generated many passionately held views. Peterson et al. (1998) have produced the methodologically strongest study. They demonstrate that reductions in mortality associated with cardiac surgery in New York State are greater than the national trend in the United States. They found no evidence of decreased access to cardiac surgery among elderly patients with acute myocardial infarction or among higher-risk elderly subsets.

Outside of cardiac surgery, few studies provide direct evidence for clinical benefits and their results are mixed (Baker et al. 2002 & 2003; Clough et al. 2002; Hibbard et al. 2005; Longo et al. 1997; Rosenthal et al. 1997). However, indirect evidence suggests that there have been clinical benefits. For example, Lee (2007) reported that the NCQA in the United States had retired the measure used to assess the use of beta blockers in patients hospitalized with acute myocardial infarction. This was because the average performance by managed care organizations participating in the HEDIS has risen from about 60% to more than 90% over the past ten years, with little variation among plans. Since this quality measure was not implemented in a controlled fashion, caution is required when drawing causal inferences about its use in public reporting systems and this dramatic improvement over time. Lee points out that no single organization (or policy) can claim credit for this success but case studies support the premise that public reporting, and the health plans' response to it, was a contributory factor. This contribution to the increased use of beta blockers after myocardial infarction must translate into lives saved. Thus, there is indirect evidence that the use of public reporting stimulates process improvements on the part of providers and that those process improvements translate into meaningful health gains for patients.

Public reporting: potential for unintended consequences

Numerous articles have discussed the potential for adverse unintended consequences resulting from the public reporting of performance data. However, the research data on this topic are relatively scant and consist mostly of surveys of how public reporting may have changed providers' practice. For example, three articles reported that cardiac surgeons in the United States thought that public reporting had made them more reluctant to operate on high-risk patients (Burack et al. 1999; Narins et al. 2005; Schneider & Epstein 1996). Similarly, Mannion et al. (2005) found that senior managers and clinicians believed that the English star performance ratings had led to a distortion of clinical priorities, erosion of public trust and reduced staff morale. However, Bridgewater et al's (2007) study in England found no evidence that public reporting had resulted in a decrease in the number of high-risk cardiac surgery cases. In fact, the proportion of high-risk cases

increased from 14.1% to 16.8% over an eight-year period in which public reporting of cardiac surgery outcomes occurred.

We have already reviewed the American data about whether or not the improvement in mortality following cardiac surgery is due to real change or to avoidance of operations for high-risk patients (Dranove et al. 2003; Moscucci et al. 2005; Omoigui et al. 1996; Peterson et al. 1998). Baker et al. (2002) reported that any benefits in in-hospital mortality rates were offset by increases in mortality post discharge in Cleveland hospitals participating in the Cleveland Health Quality Choice programme. There have been no studies of the vital issue of whether providers' attention to areas subject to public reporting comes at the expense of attention to other areas of care that may be equally or more important.

Evidence about public reporting

Public reporting has been operating in the United States for almost twenty years; perhaps unsurprisingly the source of virtually all the published data about evaluations of public reporting. However, these data concern only a small handful of the numerous public reporting systems in use.

The lack of data from other countries gives some reason to pause. If policy-makers judge that a cultural component is contributing to the effect of public reporting, then (without their own data) they must guess how the demonstrated effects in the United States might translate to their country. One conclusion seems likely to remain unchanged as the evidence suggests that public reporting of performance data has little effect on consumers' choice of providers – even in a country known for consumerism and choice in health care. It is unlikely that this result would be any different in countries with less consumerist cultures. Conversely, in countries with a greater culture of professional responsibility than the United States the public release of performance data could exert an even greater effect on providers.

Even within the United States, only a handful of public reporting systems have been subject to evaluations. Most studies consider the New York State CSRS; CAHPS; QualityCounts; California Hospital Outcomes Project; Cleveland Health Quality Choice; Pennsylvania Health Care Cost Containment Council; and HEDIS. The effects of other major reporting systems have not received peer-reviewed evaluations.

Conclusions and the challenges ahead

Our review of the literature suggests that implementation of the public reporting of quality information will stimulate providers to start or enhance activities in order to improve their performance on publicly reported measures. In Chapter 5.5, Epstein suggests a variety of criteria to consider when choosing a performance measure – strong scientific underpinning, risk adjustment for outcome measures, allow exclusions, etc. An additional criterion is required for policy-makers considering the implementation of public reporting – choose measures that assess the most important aspects of health care. This is because a measure's inclusion in a public reporting system drives the health-care system to do it and can have good effects: for example, the lives saved by near universal use of beta blockers following acute myocardial infarction or the lowering of mortality associated with cardiac surgery.

But there is also potential for negative effects. No health effect will be gained from a measure that is not linked tightly to outcomes and the resources spent might be better used on, or at the expense of, some other aspect of care. Too often the items that have been reported are those that are most expediently measured, chosen from existing datasets that will require no new data collection. Policy-makers should focus on what is important for their health-care system and aim towards a measurement system that reflects that, rather than letting the availability of existing data drive the decision about which measures will be reported.

Countries with, or considering, public reporting systems[1]

United States

The United States has numerous public reporting systems and it is not possible to list them all in this chapter. Some of the more prominent systems are described below.

HEDIS

One of the oldest and most mature public reporting systems, HEDIS is run by the NCQA (www.ncqa.org), a private not-for-profit corporation. It reports publicly on health plans that voluntarily agree the

1 Additional material from Jako Burghers

number of changes in measures from year to year. Thirty-five measures of 'the effectiveness of health care' were included in 2007.

New York State CSRS/PCI Reporting System
Oldest, best-known and most studied system for reporting short-term outcomes of cardiac interventions (www.nyhealth.gov/statistics/).

Pennsylvania Health Care Cost Containment Council – cardiac care
Another cardiac surgery system that is mature and has been the subject of research reports (www.phc4.org/reports/cabg/).

California Outcomes Reports
With a population of similar size to that of England, California is the largest American state to report some health outcomes – all at the hospital level. Outcomes are reported for cardiac surgery, community acquired pneumonia and myocardial infarction (www.oshpd.state.ca.us/HID/DataFlow/HospQuality.html).

HealthGrades
For-profit company that sells reports about doctors, hospitals and nursing homes (www.healthgrades.com/).

QualityNet
Established by the CMS, QualityNet (www.qualitynet.org/) provides the health-care quality improvement news; resources; and data reporting tools and applications used by health-care providers and others. Publicly reported quality information is made available through a companion site – Hospital Compare (see below).

Hospital Compare
Established by the CMS and members of the Hospital Quality Alliance, Hospital Compare (www.hospitalcompare.hhs.gov/) is a public-private collaboration to promote reporting on hospital quality. It displays rates for process of care measures as well as thirty-day risk adjusted mortality rates. Process measures include the antibiotic, vaccine and oxygenation status of patients with pneumonia and the provision of ACE inhibitors, aspirin and beta blockers to patients admitted for myocardial infarction; smoking cessation counselling to certain patients; and prophylactic antibiotics prior to surgery.

England

Dr Foster
Dr Foster (www.drfoster.co.uk/) is a partnership between the Health and Social Care Information Centre and Dr Foster, a private company. Its reports about Trusts in England include information about the number of operations; lengths of stay; readmission rates; nurses per 100 beds, etc., as well as hospital standardized mortality ratios.

Heart Surgery in the United Kingdom
Developed by the Care Quality Commission in collaboration with the Society for Cardiothoracic Surgery in Great Britain and Ireland and patients who have had experience of heart surgery, this web site (www.heartsurgery.cqc.org.uk/) presents risk-adjusted outcomes for cardiac surgery at thirty-nine hospitals. The EuroSCORE logistic model is used to calculate expected survival rates.

Denmark

National Indicator Project
Established in 1999, the National Indicator Project (www.nip.dk) is the result of concerted action between a number of Danish institutions, including the Ministry of Health. It measures the quality of care provided by hospitals in order to create public awareness about the extent to which health services meet quality standards. Sets of performance indicators are used to collect information on eight common conditions (stroke, hip fracture, schizophrenia, acute surgery, heart failure, lung cancer, diabetes, chronic obstructive pulmonary disease). Participation is mandatory for all hospitals. Data are published nationally, allowing benchmarking of hospitals.

Unit of Patient Evaluation
This organization has conducted a biennial survey of patients' experiences of hospital care since 2000. The data are aggregated on a national level that enables information to be used for improving hospital quality but not for hospital selection.

Other relevant organizations and web sites
Several Danish websites provide information on public and private hospitals, e.g. waiting times, treatment options, number of surgical interventions, follow-up care.

Relevant organizations and web sites include:

- Danish HealthCare Quality Programme (http://www.ikas.dk/English.aspx)
- Sundhed.dk (http://www.sundhed.dk/wps/portal/_s.155/18 6)
- Sundhedskvalitet (Health Quality) (http://www.sundhedskvalitet.dk).

Germany

Several organizations report on the quality of healthcare.

Bundesgeschäftsstelle Qualitätssicherung (BQS)
Independent organization established by the government, responsible for clinical performance assessment which is mandatory for all hospitals in Germany (used 212 indicators in 2004; 169 in 2005). Results are integrated in quality reports that include recommendations for improvement. Data on individual hospitals are not published, so consumers cannot use them for selection purposes (www.bqs-online.com).

Bertelsmann Stiftung
Conducts an annual health survey (Gesundheitsmonitor) of the experiences and needs of professionals and consumers. Since 2008, quality information has been provided through a web site (weisse-liste.de) developed and maintained by the Bertelmanns Stiftung (www.bertelsmann-stiftung.de).

Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG)
Independent scientific institute (http://www.iqwig.de) established in the course of the health care reform in 2004. Evaluates the quality and efficiency of health care and also publishes health information for patients and the general public. Primary goal is to contribute to

improvements in health care in Germany. German/English web site was launched in July 2005 as part of IQWiG's legislative remit to inform the public (http://www.Gesundheitsinformation.de). Web site includes information for consumers and patients, based on the Institute's own scientific publications and topics of its choice, but does not contain quality information on individual hospitals.

Other relevant organizations and web sites
Patienten-information (www.patienten-information.de).

Netherlands

There is increasing attention on the transparency of health-care quality in the Netherlands. Several organizations (governmental, professional and insurance companies) have developed performance indicators in many disease areas. Initially health-care practitioners and hospitals were targeted in order to encourage quality improvement and to enable benchmarking.

In 2006, a reform of the Dutch health-care system offered consumers more opportunities for choice. Health-care insurers invested heavily in promoting their plans and, as a result, 20%-30% of consumers changed their insurance plan. However, this proportion is now decreasing (no more than 5% change was expected in 2008). In 2005, the Ministry of Health, Welfare and Sport launched a web site (www.kiesbeter.nl) to provide consumers with health information and comparative information on hospital care and health insurers in order to enable better choices. The web site includes quality information and performance assessment of individual hospitals.

Private initiatives include the top 100 hospitals list produced by the daily newspaper, *Algemeen Dagblad*; the Best Hospitals list published by Elsevier; and web sites that offer comparisons of hospitals and other health-care services (e.g. www.mediquest.nl; www.independer.nl).

As in Denmark and Germany, there are no systematic data available on the effect of public reporting on the selection of health-care services, quality improvement and patient outcomes. Nevertheless, politicians and policy-makers in particular have a strong belief that quality information will result in improvements in the quality of health care and more informed decision-making among consumers.

Other relevant organizations and web sites
National Institute for Public Health and the Environment (http://www.rivm.nl).
DGN Publishers BV (private) (www.zorgkiezer.nl).

Norway

National quality indicators for the specialized health-care services were introduced in Norway in 2003. In 2006, data for twenty-one indicators were registered (11 for somatic care; 10 for psychiatric care) including patient experience surveys. The reporting of data is compulsory and they are published online (www.frittsykehusvalg.no) together with information about waiting times for different treatments and initiatives. Data are presented at an organizational (hospital) level and on the national average. Developments over time are also shown.

References

Baker, DW. Einstadter, D. Thomas, CL. Husak, SS. Gordon, NH. Cebul, RD (2002). 'Mortality trends during a program that publicly reported hospital performance.' *Medical Care*, 40(10): 879–890.

Baker, DW. Einstadter, D. Thomas, C. Husak, S. Gordon, NH. Cebul, RD (2003). 'The effect of publicly reporting hospital performance on market share and risk-adjusted mortality at high-mortality hospitals.' *Medical Care*, 41(6):729–740.

Bentley, JM. Nash, DB (1998). 'How Pennsylvania hospitals have responded to publicly released reports on coronary artery bypass graft surgery.' *Joint Commission Journal on Quality Improvement*, 24(1): 40–49.

Berwick, DM. James, B. Coye, MJ (2003). 'Connections between quality measurement and improvement.' *Medical Care*, 41(Suppl. 1): I30–38.

Bridgewater, B. Grayson. AD. Brooks, N. Grotte, G. Fabri, BM. Au, J. Hooper, T. Jones, M. Keogh B. North West Quality Improvement Programme in Cardiac Interventions (2007). 'Has the publication of cardiac surgery outcome data been associated with changes in practice in northwest England: an analysis of 25,730 patients undergoing CABG surgery under 30 surgeons over eight years.' *Heart*, 93(6): 744–748.

Burack, JH. Impellizzeri. P. Homel, P. Cunningham, JN Jr. (1999). 'Public reporting of surgical mortality: a survey of New York State cardiothoracic surgeons.' *Annals of Thoracic Surgery*, 68(4): 1195–1200; discussion: 1201–1202.

Chassin, MR (2002). 'Achieving and sustaining improved quality: lessons from New York State and cardiac surgery.' *Health Affairs (Millwood)*, 21(4): 40–51.

Clough, JD. Engler, D. Snow. R. Canuto, PE (2002). 'Lack of relationship between the Cleveland Health Quality Choice project and decreased inpatient mortality in Cleveland.' *American Journal of Medical Quality*, 17(2): 47–55.

Dranove, DEA. Kessler, D. McClellan, M. Satterthwaite, M (2003). 'Is more information better? The effects of "report cards" on health care providers.' *Journal of Political Economy*, 111(3): 555–588.

Dziuban, SW Jr. McIlduff, JB. Miller, SJ. Dal Col, RH (1994). 'How a New York cardiac surgery program uses outcomes data.' *Annals of Thoracic Surgery*, 58(6): 1871–1876.

Farley, DO. Elliott, MN. Short, PF. Damiano, P. Kanouse, DE, Hays, RD (2002). 'Effect of CAHPS performance information on health plan choices by Iowa Medicaid beneficiaries.' *Medical Care Research and Revue*, 59(3): 319–336.

Farley, DO. Short, PF. Elliott, MN. Kanouse, DE. Brown, JA. Hays, RD (2002a). 'Effects of CAHPS health plan performance information on plan choices by New Jersey Medicaid beneficiaries.' *Health Services Research*, 37(4): 985–1007.

Fung, CH. Lim, Y. Mattke, S. Damberg, C. Shekelle, PG (2008). 'Systematic review: the evidence that releasing performance data to the public improves quality of care.' *Annals of Internal Medicine*, 148(2): 111–123.

Ghali, WA. Ash, AS. Hall, RE. Moskowitz, MA (1997). 'Statewide quality improvement initiatives and mortality after cardiac surgery.' *Journal of the American Medical Association*, 277(5): 379–382.

Hannan, EL. Kilburn, H Jr. Racz, M. Shields, E. Chassin, MR (1994). 'Improving the outcomes of coronary artery bypass surgery in New York State.' *Journal of the American Medical Association,* 271(10): 761–766.

Hannan, EL. Kumar, D. Racz, M. Siu, AL. Chassin, MR (1994a). 'New York State's Cardiac Surgery Reporting System: four years later.' *Annals of Thoracic Surgery*, 58(6): 1852–1857.

Hannan, EL. Siu, AL. Kumar, D. Kilburn, H Jnr. Chassin, MR (1995). 'The decline in coronary artery bypass graft surgery mortality in New York State. The role of surgeon volume.' *Journal of the American Medical Association*, 273(3): 209–213.

Harris, KM (2002). 'Can high quality overcome consumer resistance to restricted provider access? Evidence from a health plan choice experiment.' *Health Services Research*, 37(3):551–571.

Hibbard, JH. Stockard, J. Tusler, M (2003). 'Does publicizing hospital performance stimulate quality improvement efforts?' *Health Affairs (Millwood)*, 22(2): 84–94.

Hibbard, JH. Stockard, J. Tusler, M (2005). 'Hospital performance reports: impact on quality, market share, and reputation.' *Health Affairs (Millwood)*, 24(4): 1150–1160.

Jha, AK. Epstein, AM (2006). 'The predictive accuracy of the New York State coronary artery bypass surgery report-card system.' *Health Affairs (Millwood)*, 25(3): 844–855.

Lee, TH (2007). 'Eulogy for a quality measure.' *New England Journal of Medicine*, 357(12): 1175–1177.

Longo, DR. Land, G. Schramm, W. Fraas, J. Hoskins, B. Howell, V (1997). 'Consumer reports in health care. Do they make a difference in patient care?' *Journal of the American Medical Association*, 278(19): 1579–1584.

Luce, JM. Thiel, GD. Holland, MR. Swig, L. Currin, SA. Luft, HS (1996). 'Use of risk-adjusted outcome data for quality improvement by public hospitals.' *Western Journal of Medicine*, 164(5): 410–414.

Mannion, R. Davies, H. Marshall, M (2005). 'Impact of star performance ratings in English acute hospital trusts.' *Journal of Health Services Research and Policy*, 10(1): 18–24.

Marshall, MN. Shekelle, PG. Leatherman, S. Brook, RH (2000). 'The public release of performance data: what do we expect to gain? A review of the evidence.' *Journal of the American Medical Association*, 283(14): 1866–1874.

Mennemeyer, ST. Morrisey, MA. Howard, LZ (1997). 'Death and reputation: how consumers acted upon HCFA mortality information.' *Inquiry,* 34(2):117–128.

Moscucci, M. Eagle, KA. Share, D. Smith, D. De Franco, AC. O'Donnell, M. Kline-Rogers, E. Jani, SM. Brown, DL (2005). 'Public reporting and case selection for percutaneous coronary interventions: an analysis from two large multicenter percutaneous coronary intervention databases.' *Journal of the American College of Cardiology*, 45(11): 1759–1765.

Mukamel, DB. Mushlin, AI (1998). 'Quality of care information makes a difference: an analysis of market share and price changes after publication of the New York State Cardiac Surgery Mortality Reports.' *Medical Care*, 36(7): 945–954.

Mukamel, DB. Mushlin, AI. Weimer, D. Zwanziger, J. Parker, T. Indridason, I (2000). 'Do quality report cards play a role in HMOs' contracting practices? Evidence from New York State.' *Health Services Research*, 35(1 Pt 2): 319–332.

Mukamel, DB. Weimer, DL. Zwanziger, J. Gorthy, SF. Mushlin, AI (2004–2005). 'Quality report cards, selection of cardiac surgeons, and racial disparities: a study of the publication of the New York State Cardiac Surgery Reports.' *Inquiry,* 41(4): 435–446.

Mukamel, DB. Weimer, DL. Zwanziger, J. Mushlin, AI (2002). 'Quality of cardiac surgeons and managed care contracting practices.' *Health Services Research*, 37(5): 1129–1144.

Narins, CR. Dozier, AM. Ling, FS. Zareba, W (2005). 'The influence of public reporting of outcome data on medical decision making by physicians.' *Archives of Internal Medicine*, 165(1): 83–87.

Omoigui, NA. Miller, DP. Brown, KJ. Annan, K. Cosgrove, D 3rd. Lytle, B. Loop, F. Topol, EJ (1996). 'Outmigration for coronary bypass surgery in an era of public dissemination of clinical outcomes.' *Circulation*, 93(1): 27–33.

Peterson, ED. DeLong, ER. Jollis, JG. Muhlbaier, LH. Mark, DB (1998). 'The effects of New York's bypass surgery provider profiling on access to care and patient outcomes in the elderly.' *Journal of the American College of Cardiology*, 32(4): 993–999.

Rainwater, JA. Romano, PS. Antonius, DM (1998). 'The California Hospital Outcomes Project: how useful is California's report card for quality improvement?' *Joint Commission Journal on Quality Improvement*, 24(1): 31–39.

Rosenthal, GE. Hammar, PJ. Way, LE. Shipley, SA. Doner, D. Wojtala, B. Miller, J. Harper, DL (1998). 'Using hospital performance data in quality improvement: the Cleveland Health Quality Choice experience.' *Joint Commission Journal on Quality Improvement*, 24(7): 347–360.

Rosenthal, GE. Quinn, L. Harper, DL (1997). 'Declines in hospital mortality associated with a regional initiative to measure hospital performance.' *American Journal of Medical Quality*, 12(2): 103–112.

Schneider, EC. Epstein, AM (1996). 'Influence of cardiac-surgery performance reports on referral practices and access to care. A survey of cardiovascular specialists.' *New England Journal of Medicine*, 335(4): 251–256.

Spranca, M. Kanouse, DE. Elliott, M. Short, PF. Farley, DO. Hays, RD (2000). 'Do consumer reports of health plan quality affect health plan selection?' *Health Services Research*, 35(5 Pt 1): 933–947.

Tu, JV. Cameron, C (2003). 'Impact of an acute myocardial infarction report card in Ontario, Canada.' *International Journal for Quality in Health Care*, 15(2): 131–137.

Vladeck, BC. Goodwin, EJ. Myers, LP. Sinisi, M (1988). 'Consumers and hospital use: the HCFA "death list".' *Health Affairs (Millwood)*, 7(1):122–125.

5.3 *Developing information technology capacity for performance measurement*

THOMAS D. SEQUIST, DAVID W. BATES

Introduction

Health information technology (IT) plays a substantial role in performance measurement in many locations, particularly as such measurement programmes seek to involve a broad-based collection of health systems, payers, hospitals and individual clinicians. This role should soon become even greater as information technologies (e.g. electronic health records, data warehouses, electronic claims) can provide ready access to the clinical information required to assess quality of care across a broad spectrum of conditions and among large populations.

Electronic information systems have distinct advantages over paper review and administrative data, including the standardization of data collection; provision of expanded clinical detail; and the ability to update information in real time. However, these benefits are accompanied by significant upfront and ongoing challenges such as developing the infrastructure for installing and maintaining such systems; standardizing data collection; and ensuring comparability across systems. Despite this, clinical information systems should soon become the key platform for performance measurement in developed countries and will also play a substantial role in future programmes for improving health-care quality.

This chapter explores several key issues regarding the use of IT for performance measurement, including the required infrastructure for, and penetration of, such technology; its potential capabilities; and specific issues that arise when IT is used to measure quality of care.

Infrastructure of health information network

Health IT requires a robust infrastructure if it is to be used for performance measurement. This infrastructure can be viewed at the local,

regional or national level; all with distinct yet complementary goals. Local implementation of health information networks facilitates quality measurement and reporting for a given health plan, hospital or clinic and allows the development of local initiatives to improve care and to assess their effectiveness. However, such local efforts present challenges to attempts to assess performance across settings. Comparisons can be difficult as independent health information systems may not share the same standards for data representation and are likely to have even more variable data collection methods. However, the implementation of national standards for data representation and measurement and regional and national health information networks can standardize measure reporting at the regional and local levels and allow broader assessments of clinical performance.

Infrastructure requirements at the level of local hospitals and clinics depend to some extent on the type of health information to be used for performance assessment, ranging from the use of administrative claims data to a fully functional electronic health record. The former are dependent on electronic claims submissions, requiring the establishment of computerized databases that function in the background with no real-time interaction with the live clinical environment. These data warehouses can be maintained by technical support staff and updated at intervals that fit performance measurement and quality improvement. Claims data have been convenient sources for some time but it is likely that they will be superseded by clinical data from electronic health records.

The implementation of a fully functional electronic health record entails a much larger commitment than a claims database, to both support and maintain (Poon et al. 2004). The infrastructure needs to encompass live clinical environments including patient scheduling; laboratory, radiology and pharmacy systems; and clinical notes. Background data systems are also vital as consistent and reliable data entry provides the basis for valid performance measurement. This will include certain key elements: (i) ensuring the availability of networked personal computer access in all clinical workspaces; (ii) maintaining high speed interactivity among these computers; (iii) allowing structured data entry of those fields that inform performance measurement activities; and (iv) eliminating the need and potential for data entry workarounds that will not be captured (e.g. hand-written or verbal orders). Relevant data collected in the live clinical environments can

be backed up routinely to create general data warehouses and data marts focused on particular diseases, such as a diabetes registry. Data warehouses are essential for queries across large numbers of patients, as required for quality assessment. The architecture of clinical databases is not suited to such queries which can bring operational databases to a grinding halt.

The extension of performance measurement from the local level to the regional or national level requires consideration of the involved parties; determination of a focus on hospital versus office-based care; and data storage and exchange. Ideally this will ensure comprehensive performance measurement by involving clinical providers, payers, clinical laboratories and pharmacies (Kaushal et al. 2005). A comprehensive selection of clinical providers (including hospitals, physician office practices, skilled nursing facilities, home health agencies) allows the collection of data on the full spectrum of clinical care, including patient demographics; diagnoses and procedures; medication utilization; and laboratory testing and results across hospital and office-based settings (Kaushal et al. 2005).

Performance measurement can take place in either the hospital or the office setting. However, quality assessment sometimes requires knowledge of care across both settings and the importance of transitions has been increasingly recognized. The targeted areas of performance assessment will guide the decision to focus on a particular setting for the purposes of establishing an adequate infrastructure. Some measures of care are largely hospital-based, e.g. the Hospital Quality Alliance measures on timing of antibiotic administration for treatment of pneumonia and use of aspirin for treatment of acute myocardial infarction (Jha et al. 2005). Some are focused largely on office-based care, including mammography for breast cancer screening (Trivedi et al. 2005). Others require knowledge of care in both the hospital and the office setting – for example, asthma management focuses on both medication use and the frequency of hospital visits. Once a set of measures has been identified the spectrum of required providers can be narrowed or expanded to ensure adequate data capture. The key issue for health IT is what variables need to be collected, ideally as a part of routine care. It can be especially onerous to collect some exclusion criteria and contraindications and those who develop the measures should consider whether or not they are all worthwhile.

A variety of models can be employed for data storage and exchange at the regional and national level. These might differ according to the heterogeneity of systems used to collect data; site of electronic data storage; and the strength of networking among sites. One model uses a single information system – participating organizations use one network to feed information into a central server that acts as a hub for storage and analysis. This model facilitates ready access to a completely standardized set of clinical data that allows immediate performance assessment at the national level. This creates substantial potential for uniform performance measurement but requires a system that is built from the ground up – installing the unique hardware and software at all participating clinical provider sites, for pharmacy and laboratory systems and for payer groups. In addition, the storage of data from local clinical sites on a single national server creates substantial concern about data security and the privacy of health information and necessitates the implementation of policies and procedures to safeguard such information. These policies include regulations regarding who may access the clinical data and for what specific purpose; and also to determine whether patient permission to store data outside of the local clinical site needs to be obtained prospectively. Such homogeneity is difficult to achieve and is the rare exception.

The national health information infrastructure in the United Kingdom is similar to the model described above although it does include multiple different electronic health records (Chantler et al. 2006). In 2002, the NHS began large investments in a national health information system that would facilitate widespread measurement and improvement of health-care delivery. Within the resulting national broadband network, the Spine stores demographic information on every citizen in England (including name, date of birth, address, registered primary care physician, unique patient identifier). Connected to over 98% of general practices in England, this provides a near complete listing of all patients in the country. Five regional service providers were created to direct the implementation of electronic patient records at all clinics in the country and several vendors operate electronic health records within each service area. Detailed clinical data are abstracted automatically from these records to create patient summaries of important diagnoses and procedures, laboratory results and prescriptions. Patient summary records are stored on the Spine to

allow regional and national assessments of health-care delivery. This model highlights the vast potential of a planned implementation of a national health information infrastructure. However, there are concerns about the ongoing expense of maintaining the infrastructure; shortcomings in the system's technical capacity to manage the vast amount of clinical data being generated; and the transferability of the system to new regions including Scotland and Wales.

An alternative approach would allow local organizations to implement their own technologies (around a set of data representation and exchangeability standards) and to create health information exchanges that would transfer, rather than store, clinical information. A model close to this is being developed in the United States. Under the leadership of the Office of the National Coordinator for Health Information Technology (ONCHIT), regional centres or health information exchanges will facilitate the merging of data from disparate sites to allow the combination of data within larger geographical units. This model has the advantage of allowing local health organizations to use existing systems and avoids the permanent storage of data outside the local clinical organization. However, there are also significant disadvantages – for example, difficulties with the standardization of data formats may impede data merging. In addition, data ownership ultimately resides at the local level which will need to be approached for each new performance assessment or national estimates of quality of care. One key issue is how many electronic records to include in each region – the interoperability in the United Kingdom system is due in part to the limited number of vendors in each region. This process is being implemented to a variable extent in the United States, e.g. the Massachusetts eHealth Collaborative [www.maehc.org].

The systems in the United Kingdom and the United States are examples of two conceptual models for implementing a national health IT infrastructure (Fig. 5.3.1). Other examples demonstrate variations of these concepts. Finland is a leader in the use of electronic health records: over 90% of practices use electronic records to document care and there is a strong push towards national use of e-prescription. Rather than creating a national spine for information transfer and storage, Finland has adopted a national IT roadmap to transmit health information between entities over secure commercially owned virtual private networks restricted to health-care purposes. The roadmap actively promotes the use of standardized formats to allow data

Finland
- Strong penetration of electronic health records.
- No national architecture dedicated to health-care information exchange.
- Data exchange accomplished via secure connections on commercially owned broadband network.
- Emphasis on adherence to data standards to ensure exchangeability.

Germany
- Focus on patient electronic health cards.
- Identify patients across providers and regions.
- Carry pertinent health information at discretion of patient.

United Kingdom
- Nationally owned and implemented infrastructure.
- Information Spine stores health information on all patients.
- Costly to implement but allows relatively complete capture of population health delivery.

United States
- Local development and implementation of health information technology tools, including electronic health records.
- Creation of regional health information exchanges.
- Reliance on adherence to data standards to ensure exchangeability.

Fig. 5.3.1 Conceptual models of IT infrastructure plans

Many countries have developed roadmaps for implementing a health IT infrastructure within improvement performance measurement and quality of care. These models often vary according to the underlying structure of the health-care delivery system within a country, including issues of finance and ownership.

exchange between systems. Countries such as Austria and Germany have focused efforts on electronic patient cards that protect health information but also identify patients across multiple components of the health-care system. This requires substantial initial investment in technical architecture to ensure that the card is compatible across the system. However, it also offers the promise of a true patient health record containing portable health information that can be used to improve the quality, safety and efficiency of health care.

Penetration of health IT

Widespread use of IT for performance measurement is dependent on the penetration of such technology among key stakeholder organizations including clinical providers; payers; and laboratory and pharmacy systems. The accuracy of performance reporting that relies solely on electronic data depends on all potential sources of data utilizing an electronic platform to store and transfer information. The use of paper systems by any one of these stakeholders could result in gaps in information and inaccurate estimates of health-care delivery. For example, an analysis of acute myocardial infarction care may miss vital information if pharmacy records are not available in an electronic form, e.g. use of beta blocker therapy following hospital discharge.

In addition, high rates of penetration are necessary to assure that performance estimates derived from electronic data provide an accurate reflection of population health and are not biased by reliance on data obtained from a unique subset of clinics that chose to implement IT. Early adopters may be more interested in quality measurement and improvement and thus provide performance assessments that are not representative of the entire population.

Specific information technologies show varying levels of adoption. One report estimates that the penetration of electronic claims submission is already relatively high in the United States and will approach 100% within the next two years (Kaushal et al. 2005). It is more challenging to estimate the use of electronic health records in the United States due to the lack of a uniform definition of what constitutes an electronic health record. This can range from a system that shows only laboratory results to a fully functional system that includes clinical decision support tools, computerized order entry and electronic note authoring (Friedman 2006). However, it is clear that most other industrialized nations have progressed further (Ash & Bates 2005).

The definition of an electronic health record can vary according to its need and purpose. There are two distinct types of electronic patient records in the United Kingdom: (i) those that describe care provided by a single institution; and (ii) those that describe a system that allows the exchange of electronic clinical data across settings to provide a complete, longitudinal representation of health-care delivery (Friedman 2006). However, there is no current requirement for specific functionalities beyond these general descriptions.

Several efforts to standardize the definition are underway in the United States. Based on much more specific requirements, these processes all endorse the need for electronic health records to support a reporting function. The Institute of Medicine has defined eight core functions of an electronic health record: (i) health information and data; (ii) results management; (iii) order management; (iv) decision support; (v) electronic communication and connectivity; (vi) patient support; (vii) administrative processes and reporting; and (vii) reporting and population health (Board on Health Care Services & Institute of Medicine 2003). In 2005, the federal government formed the Certification Commission for Healthcare Information Technology (www.cchit.org) to establish a certification process for IT based on minimum standards for functionality, security and interoperability. These standards will be used to certify not only electronic health records but also health networks that allow the exchange of data among hospitals and clinics. The CCHIT certified more than seventy-five outpatient records in its first year and is currently certifying inpatient records. There has been attention to ensuring that some quality measures can be addressed and a current process is attempting to define what atomic data elements will be needed but most functions are currently certified as either present or absent.

Despite the limitations inherent in defining electronic health records, some estimates of penetration increase understanding of the current status of IT and its potential for performance measurement (Fig. 5.3.2). The United Kingdom has made the most progress in creating a national health information architecture and implementing an electronic health record system. Recent estimates suggest that over 90% of general practices in England use electronic patient records (Chantler et al. 2006; Schoen et al. 2006), facilitating a rather complete picture of office-based care. There are similar adoption rates in most Scandinavian countries, Australia and New Zealand. North America lags behind – electronic health records are used by only 28% of physicians in the United States and 23% in Canada (Jha et al. 2006; Schoen et al. 2006). Adoption rates vary with the size of practices – larger practices have implemented electronic health records approximately two to three times more than smaller practices and solo physician practices (Jha et al. 2006). Furthermore, many systems tend to be focused on the collection of data in the ambulatory setting; fewer are designed to capture both hospital and office-based care (Chantler et al. 2006; Schoen et al. 2006).

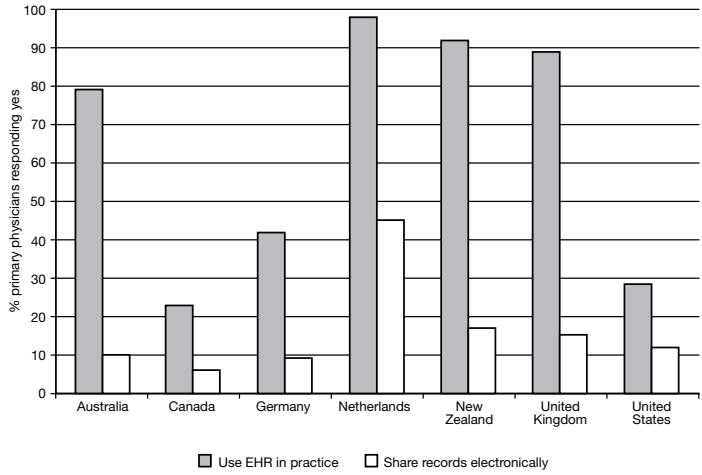


Fig. 5.3.2 International penetration of electronic health records and data exchangeability: responses from primary care physicians across seven countries, 2006

Source: Schoen et al. 2006

There are limited data regarding the use of such programmes for health information exchange at the broader regional and national levels. The United Kingdom has the most fully functioning system, storing patient level data in central repositories that allow performance reporting on a national level (Campbell et al. 2007). In the United States, the ONCHIT Nationwide Health Information Network has instituted a pilot programme in nine regional networks to investigate the feasibility of using regional health information exchanges.

Many factors affect the adoption rates for IT, particularly electronic health records, but the perception of clinical providers is paramount. It is clear that enlisting the support of management and clinicians is an essential component of successful implementation (Poon et al. 2004; Scott et al. 2005). Some high profile examples of failed implementation have resulted from clinicians' dissatisfaction with the system (Connolly 2005). Diffusion of innovations research suggests that the 'late majority' of technology adopters represent a constituency that provides the 'critical mass' necessary to ensure continued widespread use (Rogers 2003). In order to develop this critical mass, health system

leaders need to develop implementation plans that minimize upfront challenges to clinicians' workflow and efficiency and make the benefits of adoption more transparent to the general workforce.

It is equally important to understand patients' views on the adoption and use of IT. Some data suggest that patients feel that electronic health records may reduce the amount of time that their physician spends talking with them during an office visit, but very few feel that the quality of the overall interaction is diminished (Rouf et al. 2007). Innovative use of technology such as cell phones, Internet patient portals and portable electronic health records has vast potential to improve health care and patient experiences of care and to increase patients' engagement in their own health-care delivery (Smith & Barefield 2007). However, while many patients are in favour of advancing the use of IT, many also express legitimate concerns regarding the security and privacy of their health information (Chhanabhai & Holt 2007).

Capabilities of electronic health records

Having installed the IT infrastructure, it is necessary to consider electronic health records' suitability for valid assessments of performance. Performance assessment can be categorized according to the six domains identified by the Institute of Medicine to assess whether care is equal, effective, safe, efficient, patient-centred and timely (Institute of Medicine 2001). Data from electronic health records are likely to be most valuable for assessing the effectiveness, safety, efficiency and equality of health-care delivery, in both office and hospital settings.

Health-care effectiveness

Electronic health records offer clear benefits over the use of paper record reviews and administrative claims data when assessing health-care effectiveness. Paper record reviews require more personnel to identify charts for abstraction; training of chart abstracters to ensure uniformity; and manual recording of data needed for performance measurement. Given the complexity of this process, including time and personnel commitment, most performance measurement that relies on manual chart review is completed on only a limited sample of the total population.

Administrative claims data offer the substantial advantage of being available in electronic form in the vast majority of settings, thereby allowing automated identification of data for the entire population with limited expenditure. However, they offer a limited spectrum of data for useful performance measurement. Administrative claims data are intended primarily as a source for financial accounting and therefore often lack the clinical detail needed to assess important health outcomes (e.g. blood pressure control) or counselling efforts (e.g. tobacco counselling). In addition, in a multi-payer system such as that in the United States, administrative data need to be pooled across multiple payers to provide a complete performance assessment for one provider, e.g. a hospital or clinic. Electronic health record systems incur substantial capital costs (Chantler et al. 2006; Kaushal et al. 2005a) but once in place may allow performance measurement with substantially fewer resources than paper chart reviews and offer increased clinical detail that is not available in administrative claims data.

Electronic health records require several key data elements to enable reliable assessment of health-care effectiveness across a spectrum of conditions. These include patient demographics, diagnosis and procedure codes, laboratory and radiology results, pharmacy data and allergy information. All of these elements contribute to the standard assessment of quality metrics for health-care effectiveness, which includes identification of the eligible denominator and numerator populations. The creation of the eligible denominator population is reliant on all patients being assigned a unique patient identifier within the electronic health record. This is particularly important for performance measurement across multiple clinical sites in which duplicate identification of patients could threaten the validity of the analysis. In addition, metrics are often assessed by provider and this requires patients to be assigned to a specific provider, such as the primary care provider or specialist for a given condition. When unique patients have been identified and linked (if necessary) to specific providers, further eligibility criteria for the denominator population can be applied from electronic health record data. These structured fields typically include patient demographics (e.g. age and sex) and diagnostic codes (e.g. congestive heart failure or diabetes). Finally, exclusion criteria must be applied, often by using medication allergy information or other relevant data. Identification of the appropriate numerator population

from electronic health record data most often relies on laboratory and radiology results, as well as pharmacy data.

There is a growing number of examples of the use of electronic health record data to assess quality based on the principles described above (Baker et al. 2007; Benin et al. 2005; O'Toole et al. 2005; Persell et al. 2006; Tang et al. 2007). Identification of eligible denominator populations for some screening measures is straightforward and unlikely to be biased, regardless of data source. For example, quality measures constructed from electronic health record data that are strictly age-based (e.g. breast or colorectal cancer screening rates) are unlikely to include or exclude patients inappropriately. However, those measures that are based on the presence of a specific disease require increased attention to ensure that the appropriate denominator population is identified accurately.

Benin et al. (2005) assessed acute pharyngitis care by comparing electronic health record data with administrative claims data, using manual chart review as the gold standard. For identification of cases of pharyngitis they found that the electronic health record had a higher sensitivity than claims data (96% versus 62%), but a lower specificity (34% versus 55%). However, this may not provide an accurate reflection of the potential for accuracy of electronic health record data as this study identified cases through free text searches rather than coded data.

The ability to identify accurately the eligible denominator populations for chronic disease care has also been examined. In one study of Medicare patients, coded electronic health record data had substantially higher sensitivity than claims data for identification of diabetic patients (97% versus 75%), with a near perfect specificity (99.6%) (Tang et al. 2007). This high level of data accuracy was achieved primarily by using coded information in the electronic problem list; the presence of a diabetes medication on the electronic medication list; and laboratory results consistent with the presence of uncontrolled diabetes.

Electronic health record data also offer the opportunity to further refine the identification of patients with diabetes. For example, standard definitions of quality measurement for diabetes care require the presence of at least two visits for diabetes during the measurement period. This is intended to improve the specificity of the denominator population despite the fact that only 75% of patients with diabetes

meet this requirement (Tang et al. 2007). Electronic health record data can be less reliant on the number of office visits and track more patients with diabetes through electronic problem and medication lists, as well as the availability of historical laboratory data. However, diabetes is much easier to detect than many other chronic conditions (coronary artery disease, congestive heart failure, chronic obstructive pulmonary disease) since some drugs are used almost exclusively to treat diabetes and there are good laboratory markers.

There have also been assessments of the accuracy of electronic health record data for identifying appropriate numerator populations. For management of pharyngitis, electronic health record data had slightly lower rates of identified testing for Group A streptococcus than administrative data (71% versus 76%) (Benin et al. 2005). The most detailed assessment of the accuracy of numerator data comes from two studies of cardiovascular disease care in the office setting. Electronic health record data were used to evaluate standard performance measures for patients with coronary artery disease, such as measurement of cholesterol, measurement of blood pressure and use of appropriate medications (antiplatelet drugs, lipid lowering drugs, beta blockers, angiotensin converting enzyme [ACE] inhibitors). Rates of appropriate care were consistently lower when using coded electronic health record data rather than manual chart review, with absolute differences between the two methods ranging from as low as 1.8% (cholesterol control) to as high as 14.3 % (antiplatelet drug use). The high discrepancy rate for antiplatelet drug use in this study is likely due to the availability of aspirin as an over-the-counter treatment. This provides less incentive for clinicians to document prescriptions in coded format in the electronic medication list.

Similar findings are available for the assessment of quality of care for congestive heart failure in which quality metrics included assessment of left ventricular ejection fraction; use of beta blockers and ACE inhibitors; and prescription of warfarin for patients in atrial fibrillation. Again, coded electronic health record data showed lower rates of appropriate care than manual chart review data, ranging from a low of 1.9% for use of beta blocker therapy to a high of 23.2% for use of warfarin therapy (Baker et al. 2007). In contrast to the previous study, the high discrepancy rate for use of warfarin therapy among patients with atrial fibrillation was attributable to the lack of identification of

valid exclusion criteria in the electronic health record, such as a history of bleeding or mental disorder that precluded anticoagulation.

Possibly the largest scale demonstration of performance measurement based on electronic health record data originates from the United Kingdom, where the national information architecture has allowed measurement of health-care delivery across a spectrum of conditions (Campbell et al. 2007). Focused on health-care effectiveness, these measurements form the basis for a nationwide pay-for-performance programme targeting general practitioners. A remarkably high rate of quality performance has been achieved across a very large number of parameters. However, a very large amount (around 30%) of payment was based on quality. One issue emerged from providers being allowed to remove patients from the numerator and denominator for any measure – exception reporting. A few practices used this option for a very large number of their patients and the next iteration of this programme will include auditing around this issue for practices that use this option frequently.

The findings above highlight several key components in the use of electronic health records to measure the effectiveness of health-care delivery. The first is that the data contained within electronic health records can be used in a feasible manner to conduct performance assessment across a wide range of conditions. The second is that the identification of denominator and numerator populations presents challenges within the context of electronic health record data. Some assessment of validation of individual performance measures is advisable before implementing their routine use and those developing the measures should particularly consider the relative importance of specific exceptions. Finally, it is important to note that the above findings provide only an early window into the potential opportunities and pitfalls of using electronic health record data. This will require more information on the extension of these findings to other practice settings that use a range of electronic health record systems.

Patient safety

The standardized assessment of patient safety is a crucial imperative given the large body of evidence documenting the unintended consequences of medical care (Institute of Medicine 1999). Adverse events

are historically substantially underreported in hospital settings (Bates et al. 2003) as systematic identification and reporting systems have been difficult to implement. Clearly many injuries occur in other settings but data about these adverse events are even more limited and it has been suggested that the magnitude of harm outside the hospital may be as great as inside. Electronic health records have the potential to improve dramatically the measurement of patient safety across many areas.

The AHRQ has developed a set of hospital-based patient safety indicators (PSIs) that allow hospitals to assess patient safety and evaluate interventions to improve safety (Agency for Healthcare Research and Quality 2006). These rely on diagnostic codes to identify potential threats to patient safety such as the occurrence of incident decubitus ulcers or foreign bodies left during procedures. This information can be abstracted readily from inpatient electronic medical record systems but these codes for adverse events are not well-represented in the overall coding schemes and are not used consistently by clinicians. These sensitivity and specificity problems limit their potential to detect patient safety issues (Bates et al. 2003).

Additional strategies to detect threats to patient safety may be employed using electronic health records in hospital settings. Once electronic data are widely available, algorithms can be developed and validated to detect adverse medical events. For example, searches for key words in electronic discharge summaries can identify a spectrum of adverse events including falls, decubitus ulcers, postoperative complications, adverse drug events and unexpected death (Murff et al. 2003). Other elements of inpatient electronic health records can also be utilized. Pharmacy records can be searched for the use of medications (e.g. diphenhydramine, naloxone) commonly associated with adverse events (Classen et al. 1991). Laboratory records can be searched for out-of-range values associated with adverse events such as abnormal coagulation studies (Classen et al. 1991). Radiology reports can be searched to identify evaluations following patient falls, such as X-rays and head computed tomograms (Hripcsak et al. 1995). More advanced solutions have also been developed – natural language processing is used to discern patterns within unstructured data such as radiology reports – and their use is likely to increase as the software continues to advance (Bates et al. 2003). Alternatively, as more structured reporting of results is implemented (e.g. pathology and radiology reporting) the use of structured electronic health record data will become increas-

ingly relevant to the detection of adverse events. The general approach uses IT to detect signals that an adverse event might be present and follows this with further chart review. This needs refinement but is likely to represent the approach of the future.

There is no widely accepted set of patient safety indicators in the office setting but electronic health records have been used to detect adverse events. In particular, one study identified a substantial number of adverse drug events by using an ambulatory electronic health record and a variety of searching techniques including text search, allergy records and administrative billing codes (Honigman et al. 2001). This study highlighted the fact that the predominance of adverse events was detected using free text searches rather than structured data fields. However, increased attention to structured data entry and improvements in natural language processing will likely result in improved identification of adverse events in the office setting.

Electronic health records also have the potential to increase dramatically the measurement of another key aspect of patient safety. Follow-up of abnormal test results is a problem in many settings but may be a particular problem in the office setting where care is coordinated across many providers and health centres. Findings such as abnormal mammograms (McCarthy et al. 1996) and abnormal faecal occult blood tests (Etzioni et al. 2006) often lack adequate follow-up, diminishing the effectiveness of population based screening programmes. Through innovative use of laboratory and radiology data, electronic health records can be used to identify abnormal test results and measure the adequacy of follow-ups according to rigorously defined guidelines (Poon et al. 2004a).

Similarly, transitions in care from the hospital to the office setting are often cited as sources of considerable concern for patient safety (Roy et al. 2005). In this setting, data from electronic health records can be employed to identify abnormal test results and measure whether appropriate follow-up has occurred. One challenge to this use of electronic data is the availability of structured information to identify such abnormal results. Blood test results may have clear thresholds but other findings may be subtler and require clearly structured definitions.

Categorization schemes have been implemented in the clinical setting for topics such as mammogram interpretations, pap smear findings and colorectal polyp characteristics. However, automated identification is difficult as they are often entered into free text reports. Increased use

of advanced coding systems such as the Systemized Nomenclature of Medicine (SNOMED) algorithm will provide the structured fields that will help to solve this issue (College of American Pathologists 1984).

Health-care efficiency

Currently, there is no widely accepted set of metrics for health-care efficiency although electronic health records present an opportunity to increase measurement in this area. According to the Institute of Medicine (2001), inefficiency in health-care systems is a product of quality waste and administrative costs. Quality waste includes redundant test ordering (often due to a lack of access to prior clinical information) as well as inappropriate test use (e.g. routine use of imaging for lower back pain). Prior analyses have indicated that repeat laboratory testing in the absence of a clinical indication accounts for up to 30% of all utilization and is a particular problem in the hospital setting (van Walraven & Raymond 2003). Electronic health records hold particular promise for the measurement of such quality waste within health-care systems – they provide ready access to data on laboratory test utilization and can be used to measure rates of redundant test ordering in a reliable manner (Bates et al. 1998 & 1999).

There is increasing attention on the development of more robust measurement of health-care efficiency in both hospital and office settings. The episode treatment group is one potential option, focusing on the longitudinal management of specific conditions in both settings (Forthman et al. 2000). This technique requires access to a combination of hospital-, ambulatory- and pharmacy-based information to represent accurately the management of a specific condition, such as chronic sinusitis. Episode treatment groups can be used to identify variation in the use of procedures and medications as well as repeat office visits. Electronic health records with complete integration between hospital and office settings provide ready access to the required data and thus offer the potential to use such methodologies to assist in the measurement of efficiency.

Health-care equality

Inappropriate differences in the quality of health care are widespread throughout many health-care systems in the world and disadvantaged

populations often receive poorer quality care (Institute of Medicine 2002). These differences are based on patient socio-demographic features including sex, race, income and educational attainment. Reliable and routine measurement of such differences in care represents an important first step in the development of programmes to ensure the delivery of equal treatment to all patients. This requires the use of health information systems that can not only produce reliable data on standard measures of clinical performance but also combine these with patient level socio-demographic features. Patient gender is routinely available in administrative claims data, allowing an analysis of gender differences in health care (Ayanian & Epstein et al. 1991). However, data on patient race, income and educational attainment are far less complete (Nerenz & Currier 2004). Analyses of racial disparities in health care are often limited to black-white differences in care due to a lack of data on other racial and ethnic groups (Sequist & Schneider 2006). Similarly, patient-level income and educational attainment are often estimated at larger geographical levels, despite the known limitations of these estimates (Krieger et al. 2003).

Electronic health records can provide a reliable means of measuring health disparities according to a wide range of patient socio-demographic features (Sequist et al. 2006). Patient information (including patient race and educational attainment) can be collected as part of routine care and combined with clinical data to construct stratified measures of health-care quality.

Key issues concerning use of electronic data

Electronic health record data have the potential to improve dramatically performance measurement as outlined above. However, this potential will not be realized without careful consideration of the key issues of data quality and patient privacy.

Data quality

Electronic health records offer access to increased clinical detail for use in performance measurement. However, it is important to understand the accuracy of these data before using them for high stakes reporting, such as pay for performance or public reporting efforts. Some work has highlighted potential limitations in the use of electronic

medical record data for performance measurement (Baker et al. 2007; Persell et al. 2006). These limitations relate largely to two main issues. The first is that populations can be highly mobile, shifting care between physician practices within and across geographical regions. This creates challenges for complete capture of clinical care for the purposes of performance measurement, particularly when measurements rely on care delivered over a continuum. For example, accurate assessment of colorectal cancer screening rates requires knowledge of the performance of colonoscopy within the previous ten years. This presents a significant challenge when patients are quite likely to have relocated or changed health-care providers during this extended timeframe. A critical solution to this issue is to facilitate electronic data exchange among health-care systems or to institute shared data warehouses that allow complete capture of clinical care processes.

The second issue is that data are entered into electronic health records for the primary purpose of routine clinical care rather than performance measurement. This may lead to deficiencies in documentation or lack of use of the structured data fields required for reporting in lieu of more convenient free text documentation of care. Similarly, exclusions that apply to specific performance measures may not be coded routinely in electronic health records, either through technical limitations or because clinicians are not aware of the need to enter such structured documentation. If it is clear that specific exclusions are important it is possible to create coded fields and stress the importance of documentation to clinicians. A recent study analysed the use of electronic health record data to assess quality of care for coronary artery disease in the office setting. This revealed that 15% to 81% of cases deemed to have failed to achieve the quality metric were found on manual chart review to have met either the quality metric or valid exclusion criteria (Persell et al. 2006). The study identified three important causal factors for both numerator and denominator inconsistencies. First, clinicians often used the diagnosis of coronary artery disease inappropriately, frequently when they were ordering tests to exclude this condition (though current reimbursement models sometimes reward this approach). Second, data were often entered in non-structured data fields – such as noting aspirin use in free text rather than the formal electronic medication list. Third, valid exclusions were not captured in structured data fields, including concepts such as patient preference and adverse medication effects.

Data collected via electronic health records are primarily for the purpose of clinical care and will certainly lack data required for broad-based performance assessment. This is a significant problem but one that can be expected to improve with time, especially if clinicians are made more aware. Missing data are less likely to be a significant concern for laboratory- or radiology-based measures that already form part of routine clinical care. The completeness of the data may be of particular relevance when assessing performance measures focused on patient education or counselling, or those in which patient refusal may play a large role. Busy clinicians typically document this type of information in unstructured notes rather than in the coded fields that allow automated performance assessments. The use of coded fields can be increased through the effective design of electronic health records that encourage their use in the context of streamlined clinical workflows and through training and performance feedback on their use (Porcheret et al. 2004). It is crucial to demonstrate a clinical 'return on investment', such as basing clinical decision support tools (electronic reminders) or performance feedback reports on data entered in these coded fields (Friedman 2006). Finally, it is important to discourage the entry of free text diagnoses by ensuring that coded fields cover the full spectrum of clinical care through the use of advanced coding systems such as the SNOMED algorithm (College of American Pathologists 1984).

When encouraging the use of coded fields, it is important to consider the special case of behaviour counselling, such as smoking status. These fields should include a 'not assessed' option set as the default response in all records to avoid the pitfall of erroneously assigning a smoking status to a patient in whom such behaviour has not actually been assessed. This will allow differentiation between those patients whose smoking status has been assessed and those with missing data.

Potential solutions can be implemented to improve data quality from electronic health records and other information systems but still it is critical to ensure the reliability of the data via routine audit or other quality assurance means. This is particularly important if data are to be used for high stakes purposes such as public reporting or pay for performance. There are relatively straightforward options for ensuring data reliability, such as crosschecking data from multiple sources. For example, discrepancies arising from comparisons of administrative claims data and the electronic health record can be examined further by a more labour intensive manual chart review of a small subset

of patients. More complex options would include random chart audits conducted by trained staff. Clearly, these are more labour intensive but may be necessary initially as performance measurement programmes that are reliant on information from new electronic systems become more widespread.

Patient privacy

As electronic health record data become increasingly available and are able to provide a greater level of clinical detail on large populations of patients, it will become more important to protect the privacy of such information when assessing performance. At the local level, safeguards need to be established to ensure that passwords and network security limit access to patient information to approved personnel only and that audit trails verify individual access. Where health information is exchanged across health systems with the ultimate goal of aggregation at the regional or national level there will need to be consideration of what type of patient information can be transmitted securely and how to transmit it. This will require a careful balance between the protection of privacy and the collection and transmission of data with enough detail to allow clinically meaningful performance assessment. The level of security of electronic transmissions should be commensurate with the level of detail contained in the data, employing encryption techniques when necessary.

The United Kingdom has been at the forefront of issues related to the privacy of patient health information and data sharing outside of the local sites for performance measurement purposes (Chantler et al. 2006). As outlined earlier, the Spine stores basic demographic information on all citizens in England. These data can be augmented in a personal summary record that contains more detailed information regarding clinical diagnoses and treatments, including prescriptions, procedures and hospital discharge summaries. The demographic information stored on the Spine is compulsory for all patients but they can dictate to what extent, if at all, more detailed information is available in the personal summary record. Access to patient medical records is monitored in order to ensure data security – smart cards identify health professionals as they access information and maintained audit trails detail access to each record. However, these safeguards are not

perfect. A junior official recently extracted banking data on 25 million people in the United Kingdom. The data were saved on two disks (with little protection), mailed and subsequently lost. Such incidents have generated understandable concern.

The United States has enacted the Health Insurance Portability and Accountability Act (HIPAA) which in part regulates the use of private health information. This has implications for the use of data to measure the delivery of health care (Kamoie & Hodge 2004). ONCHIT is actively considering the options for data security and patient privacy – security is one of the core components of successful certification of electronic heath record systems through this office (http://healthit.hhs.gov/portal/server.pt). The criteria required for certification under this system are similar to those in the United Kingdom, including the requirement of secure monitored access to electronic health record information as well as maintenance of a complete audit trail of access to these records.

Key policy issues

A number of lessons emerge for policy-makers in developed nations. One clear immediate priority is to create international agreement on key quality metrics. The European Union has begun this but has not yet reached broad agreement for most metrics. This creates significant challenges and resultant unnecessary incremental work when performing international comparisons.

It is clear that financial incentives are powerful motivators for the adoption of electronic health records in the outpatient sector. Some countries (e.g. United Kingdom) have achieved near universal adoption of electronic health records in general practice by paying for these systems; other countries (e.g. Australia) have used incentives-based approaches to achieve high levels of implementation. The United Kingdom has also been extremely successful with performance measurement by offering large financial incentives based on providers' performance on quality metrics extracted from the electronic record. The requirement that providers bill electronically can also provide an important incentive. Such incentive programmes have enabled a large proportion of Europe to achieve high levels of adoption of electronic health records in the outpatient setting. The current challenge is to

improve the records so as to deliver better decision support; allow providers to work efficiently; and ensure that the records can be used readily to measure performance and improve care.

Less evidence is available about how best to achieve high levels of adoption and how to use records to measure performance in the inpatient setting. In most nations, levels of implementation in inpatient facilities lag behind what is available in the outpatient sector. This requires better incentives to encourage institutions to adopt this technology and additional approaches for routine measurement of the quality of inpatient care in order to align incentives with high quality. A clear note of caution is required as financial incentives can be a double-edged sword and may promote undesirable behaviour. Incentive programmes should be viewed from a variety of perspectives and include consideration of the possibility of gaming. The effect on the quality of care should be monitored as closely in areas that are not reliant on incentives as in those that are.

Low-income countries have far less experience of how best to proceed although some transitional countries such as Brazil have achieved notable success, particularly around the larger population centres. Furthermore, it appears likely that health IT will be useful even in very low-resource environments such as Kenya (Siika et al. 2005). More research is urgently needed on how best to increase adoption of electronic health records in the inpatient setting; to address the benefits of clinical data exchange; and to identify which solutions will be most beneficial in transitional and developing nations in particular. Further evaluation of decision support and the relative costs and benefits of implementation is needed in all settings. Furthermore, research is needed to identify quality metrics which can be implemented directly through electronic records.

The future

Looking ahead, we believe that electronic health records and patient computing will remain the key technologies for measuring and improving quality. Patient computing is likely to mature within the next ten to twenty years and patients will likely begin to manage much more of their care with the assistance of health IT (Delbanco & Sands 2004).

The influence of electronic health records in the inpatient and outpatient setting will also continue to expand – with a focus on computerized provider order entry, clinical data exchange and clinical decision support.

The use of electronic health records in the ambulatory setting is essential for capturing the health of populations and moving to the next level of performance measurement. The latter should be possible in developed nations within the next five to ten years. Most actions in the inpatient setting occur as the result of an order and electronic health records should prompt providers on appropriate actions, simultaneously improving quality and facilitating performance measurement.

Today, it is technically feasible to implement widespread clinical data exchange. It should be possible to obtain a much more comprehensive picture of quality within the next ten years, once the political and social obstacles to data exchange have been overcome. Pilot programmes are required in order to determine how best to implement data exchange in a manner that does not encroach on patient privacy. Clinical decision support is one of the keys to truly dramatic improvement. Decision support is often single-synapse but could be much more sophisticated. A number of challenges for reaching the next level have been put forward recently, including: prioritizing recommendations for presentation to providers; using free text information to create recommendations; and combining decision support recommendations for patients with multiple co-morbid conditions (Sittig et al. 2008).

Conclusions

Health information technologies, particularly electronic health records, have enormous potential to increase performance measurement in a variety of areas as outlined. The ability to achieve ready access to detailed clinical information on a spectrum of conditions with minimal resource utilization is an appealing alternative to the current system of labour-intensive manual chart reviews and increasingly unsuitable administrative claims data. Used effectively, electronic health record systems can provide real-time, clinically relevant measures of health-care delivery. This potential is yet to be realized in most health-care settings as additional work is required to overcome the substantial challenges that still exist.

Challenge 1: increase penetration of electronic health records

As discussed earlier, widespread use of electronic health records is essential to ensure the validity of performance measurement comparisons across health-care settings. Some countries have very high adoption rates but others are lagging behind, particularly among small and solo physician practices. In addition, implementations in hospitals generally lag behind office practices. Policy solutions are needed to increase the adoption of electronic health records in these settings. The need for central leadership to support adoption has been highlighted repeatedly (Poon et al. 2004). In addition, successful implementation depends on minimizing the impact on clinician workflow and efficiency, with clear demonstrations of potential care improvements. Financial barriers must be overcome (Bates 2005) and better alignment of financial incentives is needed, e.g. increased reimbursements based on the presence of electronic health records or use of key functionalities such as computerized order entry.

Challenge 2: ensure data exchangeability

Successful performance measurement and the delivery of good clinical care depend on the ability to merge data from multiple systems (including pharmacy, radiology, laboratory) into a single electronic health record (McDonald 1997). If these data exist in isolation, rather than as part of a uniform clinical record, this not only increases the complexity of performance measurement but also discourages further adoption of electronic health records.

Similarly, successful coordination to produce a single electronic health record will require further efforts to ensure that this health information can be exchanged as part of a compatible regional and national health information system. This will allow performance measurement at levels that extend from the local clinical site to international comparisons. The United Kingdom has implemented what is arguably the most successful model to date, although many difficulties remain; the United States is in the process of testing a model of regional health information exchanges (Adler-Milstein et al. 2008).

Challenge 3: increase reliability of electronic health record data

Preliminary studies indicate that electronic health record data can be used for performance measurement, but the accuracy of these data varies according to the metric. Sources of inaccuracy are related to the variable entry of data into structured fields and to the lack of complete data capture across health-care settings. Efforts to improve the use of structured fields within electronic health records should focus on increasing their visibility as part of the standard clinical workflow and on providing direct benefits of the collection of such information, such as using these data to drive electronic clinical decision support tools. Improved capture of data across health-care settings will involve ensuring that all possible key stakeholders have deployed electronic data systems. This will include hospital- and office-based providers as well as pharmacy and laboratory systems. Electronic gaps in any of these systems will challenge the validity of performance assessment based on electronic health record data.

When these challenges have been addressed, health IT can realize its true potential to advance the field of performance measurement. This will facilitate widespread assessments of health-care delivery and ultimately improve the health status of the population.

References

Adler-Milstein, J. McAfee, AP. Bates, DW. Jha, AK (2008). 'The state of regional health information organizations: current activities and financing.' *Health Affairs (Millwood),* 27(1): 60–69.

AHRQ (2006). *Patient safety indicators overview*. Rockville, MD: Agency for Healthcare Research and Quality (http://www.qualityindicators.ahrq.gov/psi_overview.htm).

Ash, JS. Bates, DW (2005). 'Factors and forces affecting EHR system adoption: report of a 2004 ACMI discussion.' *Journal of the American Medical Informatics Association,* 12(1): 8–12.

Ayanian, JZ. Epstein, AM (1991). 'Differences in the use of procedures between women and men hospitalized for coronary heart disease.' *New England Journal of Medicine,* 325(4): 221–225.

Baker, DW. Persell, SD. Thompson, JA. Soman, NS. Burgner, KM. Liss, D. Kmetik, KS (2007). 'Automated review of electronic health records

to assess quality of care for outpatients with heart failure.' *Annals of Internal Medicine,* 146(4): 270–277.

Bates, DW (2005). 'Physicians and ambulatory electronic health records.' *Health Affairs (Millwood),* 24(5): 1180–1189.

Bates, DW. Boyle, DL. Rittenberg, E. Kuperman, GJ. Ma'Luf, N. Menkin, V. Winkelman, JW. Tanasijevic, MJ (1998). 'What proportion of common diagnostic tests appear redundant?' *American Journal of Medicine,* 104(4): 361–368.

Bates, DW. Evans, RS. Murff, H. Stetson, PD. Pizziferri, L. Hripcsak, G (2003). 'Detecting adverse events using information technology.' *Journal of the American Medical Informatics Association,* 10(2): 115–128.

Bates, DW. Kuperman, GJ. Rittenberg, E. Teich, JM. Fiskio, J. Ma'Luf, N. Onderdonk, A. Wybenga, D. Winkelman, J. Brennan, TA. Komaroff, AL. Tanasijevic, M (1999). 'A randomized trial of a computer-based intervention to reduce utilization of redundant laboratory tests.' *American Journal of Medicine,* 106(2): 144–150.

Benin, AL. Vitkauskas, G. Thornquist, E. Shapiro, ED. Concato, J. Aslan, M. Krumholz, HM (2005). 'Validity of using an electronic medical record for assessing quality of care in an outpatient setting.' *Medical Care,* 43(7): 691–698.

Board on Health Care Services and Institute of Medicine (2003). *Key capabilities of an electronic health record system: letter report.* Washington, DC: National Academies Press.

Campbell, S. Reeves, D. Kontopantelis, E. Middleton, E. Sibbald, B. Roland, M (2007). 'Quality of primary care in England with the introduction of pay for performance.' *New England Journal of Medicine,* 357(2): 181–190.

Chantler, C. Clarke, T. Granger, R (2006). 'Information technology in the English National Health Service.' *Journal of the American Medical Association,* 296(18): 2255–2258.

Chhanabhai, P. Holt, A (2007). 'Consumers are ready to accept the transition to online and electronic records if they can be assured of the security measures.' *Medscape General Medicine,* 9(1): 8.

Classen, DC. Pestotnik, SL. Evans, RS. Burke, JP (1991). 'Computerized surveillance of adverse drug events in hospital patients.' *Journal of the American Medical Association,* 266(20): 2847–2851.

College of American Pathologists (1984). *Systemized nomenclature of medicine, second edition*. Skokie. IL: Vol. 1 & 2.

Connolly, C (2005). 'Cedars-Sinai doctors cling to pen and paper.' *Washington Post*, 21 March 2005: p.A01.

Delbanco, T. Sands, DZ (2004). 'Electrons in flight – e-mail between doctors and patients.' *New England Journal of Medicine,* 350(17):1705–1707.

Etzioni, DA. Yano, EM. Rubenstein, LV. et al (2006). 'Measuring the quality of colorectal cancer screening: the importance of follow-up.' *Diseases of the Colon and Rectum,* 49(7): 1002–1010.

Forthman, MT. Dove, HG. Wooster, LD (2000). 'Episode Treatment Groups (ETGs): a patient classification system for measuring outcomes performance by episode of illness.' *Topics in Health Information Management,* 21(2): 51–61.

Friedman, DJ (2006). 'Assessing the potential of national strategies for electronic health records for population health monitoring and research.' *Vital and Health Statistics Series 2,* (143): 1–83.

Honigman, B. Lee, J. Rothschild, J. Light, P. Pulling, RM. Yu, T. Bates, DW (2001). 'Using computerized data to identify adverse drug events in outpatients.' *Journal of the American Medical Information Association,* 8(3): 254–266.

Hripcsak, G. Friedman, C. Alderson, PO. DuMouchel, W. Johnson, SB. Clayton, PD (1995). 'Unlocking clinical data from narrative reports: a study of natural language processing.' *Annals of Internal Medicine,* 122(9): 681–688.

Institute of Medicine (1999). *To err is human: building a safer health system*. Washington, DC: National Academies Press.

Institute of Medicine (2001). *Crossing the quality chasm. A new health system for the 21st century*. Washington, DC: National Academies Press.

Institute of Medicine (2002). *Unequal treatment. Confronting racial and ethnic disparities in health care*. Washington, DC: National Academies Press.

Jha, AK. Ferris, TG. Donelan, K. DesRoches, C. Shields, A. Rosenbaum, S. Blumenthal, D (2006). 'How common are electronic health records in the United States? A summary of the evidence.' *Health Affairs (Millwood),* 25(6): w496–507.

Jha, AK. Li, Z. Orav, EJ. Epstein, AM (2005). 'Care in U.S. hospitals – the Hospital Quality Alliance program.' *New England Journal of Medicine,* 353(3): 265–274.

Kamoie, B. Hodge, JG Jr (2004). 'HIPAA's implications for public health policy and practice: guidance from the CDC.' *Public Health Reports,* 119(2): 216–219.

Kaushal, R. Bates, DW. Poon, EG. Jha, AK. Blumenthal, D (2005). 'Functional gaps in attaining a national health information network.' *Health Affairs (Millwood),* 24(5): 1281–1289.

Kaushal, R. Blumenthal, D. Poon, EG. Jha, AK. Franz, C. Middleton, B. Glaser, J. Kuperman, G. Christino, M. Fernandopulle, R. Newhouse, JP. Bates, DW (2005a). 'The costs of a national health information network.' *Annals of Internal Medicine,* 143(3): 165–173.

Krieger, N. Chen, JT. Waterman, PD. Rehkopf, DH. Subramanian, SV (2003). 'Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures – the public health disparities geocoding project.' *American Journal of Public Health,* 93(10): 1655–1671.

McCarthy, BD. Yood, MU. Boohaker, EA. Ward, RE. Rebner, M. Johnson, CC (1996). 'Inadequate follow-up of abnormal mammograms.' *American Journal of Preventive Medicine,* 12(4): 282–288.

McDonald, CJ (1997). 'The barriers to electronic medical record systems and how to overcome them.' *Journal of the American Medical Informatics Association,* 4(3): 213–221.

Murff, HJ. Forster, AJ. Peterson, JF. Fiskio, JM. Heiman, HL. Bates, DW (2003). 'Electronically screening discharge summaries for adverse medical events.' *Journal of the American Medical Informatics Association,* 10(4): 339–350.

Nerenz, DR. Currier, C (2004). Collection of data on race/ethnicity by health plans, hospitals and medical groups. In: Ver Ploeg, M. Perrin, E (eds.). *Eliminating health disparities: measurement and data needs*. Washington, DC: National Academies Press.

O'Toole, MF. Kmetik, KS. Bossley, H. Cahill, JM. Kotsos, TP. Schwamberger, PA. Bufalino, VJ (2005). 'Electronic health record systems: the vehicle for implementing performance measures.' *American Heart Hospital Journal,* 3(2): 88–93.

Persell, SD. Wright, JM. Thompson, JA. Kmetik, KS. Baker, DW (2006). 'Assessing the validity of national quality measures for coronary artery disease using an electronic health record.' *Archives of Internal Medicine,* 166(20): 2272–2277.

Poon, EG. Blumenthal, D. Jaggi, T. Honour, MM. Bates, DW. Kaushal, R (2004). 'Overcoming barriers to adopting and implementing computerized physician order entry systems in U.S. hospitals.' *Health Affairs (Millwood),* 23(4): 184–190.

Poon, EG. Gandhi, TK. Sequist, TD. Murff, HJ. Karson, AS. Bates, DW (2004a). '"I wish I had seen this test result earlier!" Dissatisfaction with test result management systems in primary care.' *Archives of Internal Medicine,* 164(20): 2223–2228.

Porcheret, M. Hughes, R. Evans, D. Jordan, K. Whitehurst, T. Ogden, H. Croft, P. (2004). 'Data quality of general practice electronic health records: the impact of a program of assessments, feedback, and training.' *Journal of the American Medical Informatics Association,* 11(1): 78–86.

Rogers, EM (2003). *Diffusion of innovations, fifth edition*. New York: Free Press.

Rouf, E. Whittle, J. Lu, N. Schwartz, MD (2007). 'Computers in the exam room: differences in physician-patient interaction may be due to physician experience.' *Journal of General Internal Medicine,* 22(1): 43–48.

Roy, CL. Poon, EG. Karson, AS. Ladak-Merchant, Z. Johnson, RE. Maviglia, SM. Gandhi, TK (2005). 'Patient safety concerns arising from test results that return after hospital discharge.' *Annals of Internal Medicine,* 143(2): 121–128.

Schoen, C. Osborn, R. Huynh, PT. Doty, M. Peugh, J. Zapert, K (2006). 'On the front lines of care: primary care doctors' office systems, experiences, and views in seven countries.' *Health Affairs (Millwood),* 25(6): 555–571.

Scott, JT. Rundall, TG. Vogt, TM. Hsu, J (2005). 'Kaiser Permanente's experience of implementing an electronic medical record: a qualitative study.' *British Medical Journal,* 331(7528): 1313–1316.

Sequist, TD. Schneider, EC (2006). 'Addressing racial and ethnic disparities in health care: using federal data to support local programs to eliminate disparities.' *Health Services Research,* 41(4 Pt 1): 1451–1468.

Sequist, TD. Adams, A. Zhang, F. Ross-Degnan, D. Ayanian, JZ (2006). 'Effect of quality improvement on racial disparities in diabetes care.' *Archives of Internal Medicine,* 166(6): 675–681.

Siika, AM. Rotich, JK. Simiyu, CJ. Kigotho, EM. Smith, FE. Sidle, JE. Wool-Kaloustian, K. Kimaiyo, SN. Nyandiko, WM. Hannan, TJ. Tierney, WM (2005). 'An electronic medical record system for ambulatory care of HIV-infected patients in Kenya.' *International Journal of Medical Informatics,* 74(5): 345–355.

Sittig, DF. Wright, A. Osheroff, JA. Middleton, B. Teich, JM. Ash, JS. Campbell, E. Bates, DW (2008). 'Grand challenges in clinical decision support.' *Journal of Biomedical Informatics,* 41(2): 387–392.

Smith, SP. Barefield, AC (2007). 'Patients meet technology: the newest in patient-centered care initiatives.' *The Health Care Manager (Frederick),* 26(4): 354–362.

Tang, PC. Ralston, M. Arrigotti, MF. Qureshi, L. Graham, J (2007). 'Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures.' *Journal of the American Medical Informatics Association,* 14(1): 10–15.

Trivedi, AN. Zaslavsky, AM. Schneider, EC. Ayanian, JZ (2005). 'Trends in the quality of care and racial disparities in Medicare managed care.' *New England Journal of Medicine*, 353(7): 692–700.

van Walraven, C. Raymond, M (2003). 'Population-based study of repeat laboratory testing.' *Clinical Chemistry,* 49(12): 1997–2005.

5.4 *Incentives for health-care performance improvement*

DOUGLAS A. CONRAD

Introduction

In March 2007, there were approximately 148 pay-for-performance programmes in the United States (The Leapfrog Group and MedVantage ® 2007). This marked increase (from thirty-nine in 2003) reflects the growing concern to seek increased value from the expenditures of health plans and organized health-care purchasers (predominantly government, private employers, unions, consumer groups, multiple-employer trusts). The General Medical Services Contract introduced in 2004 radically transformed the NHS in England, introducing 146 quality indicators to measure primary care team performance and encompassing 10 chronic conditions, care organization and patient experience. This new set of quality performance incentives offered general practice partnerships the potential to increase their annual income by as much as 25% (Roland 2004). Similarly, policymakers in continental Europe are moving toward strategic purchasing which optimizes population health through service mix, contract design, payment systems and choice of health care (Figueras et al. 2005).

When designing appropriate performance incentives, decision-makers must incorporate the varying socio-demographic, political, economic, cultural and organizational conditions that prevail in local, regional and national environments. The incentive options available in different polities and markets largely mirror the nature of funding and health-care delivery in those areas. Particulars of policy and practice are not only influenced substantially by specific circumstances but also (at any point in time) are somewhat 'path-dependent'– shaped by history (Figueras et al. 2005). Initial conditions are important.

This book examines multiple dimensions of health system performance: population health; financial protection; individual health outcomes; clinical quality and appropriateness; responsiveness; equity;

and health system productivity. Incentives are one type of policy instrument for improving performance and inevitably confront trade-offs among these objectives. For example, improvements in clinical quality and appropriateness and individual health outcomes might be accompanied by increased cost. Similarly, improvements in the efficiency of financial protection (e.g. risk-rating health insurance premiums) may compromise financial protection for high-risk population groups and raise questions of equity.

The theoretical framework predicts how distinct incentives will impact on health system performance. The empirical evidence review emphasizes the effects on cost and quality because incentives are generally targeted most directly to those two dimensions of performance.

Definitions and distinctions

Incentives can be conceptualized as reinforcers, stimuli or catalysts of behaviour. This chapter differentiates between incentives and other mechanisms designed to influence behaviour – measurement, information, reporting, rules, constraints and organizational structures. These interact with incentives but are not incentives per se. For example, performance measurement logically must precede application of a performance incentive but should not be confused with the incentive itself. Similarly, external (public) and internal performnce reports (e.g. peer comparisons within medical groups) may induce physicians to change behaviour in response to potential doctor-switching among patients in local markets or internal competition. Behaviour change can be motivated by the possible gain or loss in self-perceived or external reputation resulting from performance reports but the actual behavioural stimulus is the indirect dollar gain (or loss) from patient-switching or the internal psychological gain (or loss) associated with a change in reputation.

Theoretical framework

Incentives vary along several margins:

- nature of incentive (reward versus penalty)
- target entity (group or individual; provider or consumer)
- type (financial or non-financial, general versus selective)

- extrinsic versus intrinsic
- behaviour subject to incentive
- magnitude
- certainty of application (ex ante versus ex post)
- frequency and duration (short- versus long-term)
- base of comparison (relative versus absolute performance).

Nature of incentives: rewards versus penalties

In classic expected utility theory (Arrow 1963; von Neumann & Morgenstern 1944) risk-averse individuals will purchase insurance at above actuarially fair prices (i.e. when the premium reflects expected losses due to the risky event) – the excess premium reflecting risk aversion. An expected penalty will trigger a larger behavioural response (loss avoidance) than a reward of equal magnitude (gain-seeking).

Kahneman and Tversky (1979) found that the certainty effect (relative over-weighting of certain prospects compared to uncertain ones) drives decision-makers to weigh losses more heavily than similar size gains. Tversky and Kahneman (1986) noted that decision-makers tend to ignore common outcomes across prospects and focus on incremental gains or losses relative to their reference point. Incentive theory provides two important lessons: (i) penalties may be more powerful stimuli than rewards; (ii) the same decision-maker will gamble or seek insurance according to his/her initial point of reference and assessment of the probabilities and magnitudes of gain or loss.

Target entity

Other things being equal, a group-level incentive payment is a less powerful motivator of individual-level behaviour change than an individual-level incentive of identical expected amount for each individual. Individual agents tend to coast on the expected efforts of others unless there is an active monitoring or disciplinary mechanism. Accordingly, if the principal (e.g. medical group practice owner) wishes individual agents (e.g. physician employees or other owners) to perform well, individual incentives such as high-powered compensation tied to individual performance are critical to success (cf. Conrad et al. 2002; Gaynor & Gertler 1995). Gaynor and Gertler's work on physician productivity shows that physicians in larger groups are more respon-

sive to high-powered (individual production-based) compensation, as might be expected if smaller groups are inherently more able to use informal monitoring and peer pressure to enforce productivity norms.

In contrast, group or team incentives are expected to induce better performance for tasks that require cooperation and coordination among individuals. Group incentives also dominate individual incentives when the desired behaviour involves organization-level structural change, e.g. adoption of IT, chronic disease registries or electronic health records. The basic principle is to incentivize at the level of the entity responsible for a given action and which stands to capture most directly the benefits and costs.

Types of incentive

There are two general types of incentives: financial and non-financial. In turn, these can be either general or selective. For example, 'pure' forms of health plan payment to medical practices (Gosden et al. 2000) – capitation, per episode of care/case, fee for service – are general, indirect financial incentives as they do not target a particular behaviour (cost per unit of service, service volume or clinical quality). Under pure capitation, physicians bear the full cost of services for each enrollee but receive no incremental dollars per service. Capitation thus encourages the lowest level of service volume per enrollee of all plan payment types. On the same reasoning, payment per case (e.g. hospital DRG rates) and per episode (package pricing for pre- and post-surgical care) induce somewhat higher levels of service; fee for service induces the highest (Conrad & Christianson 2004).

Individual physician compensation has subtler impacts within provider organizations. For example, the salaried non-owner physician will not directly realize the marginal revenues or marginal costs of his/her treatment decisions. With fee for service, non-owner physicians will directly capture those individual marginal revenues and their pro rata or individual share of marginal costs; owner-physicians will perceive the same marginal revenue incentive and even stronger marginal cost incentive. A fixed salary might create even stronger volume incentives if the salaried physician attaches a sufficiently high decision weight to marginal health benefits delivered to patients versus marginal profit per unit under fee for service. This matches Kralewski et al's (2000) findings in medical group practices.

The ownership status of the individual physician also affects economic incentives. The owner is a *residual claimant* of practice net income, i.e. captures a share of after-tax, after-cost practice returns (Fama & Jensen 1998). Thus, independent of the general compensation method (salary, fee for service or some hybrid), owners will perceive a high-powered individual incentive to manage revenues and costs. Moreover, ownership confers a non-financial, *reputational* incentive to take actions that will enhance the brand name of the practice, such as the optimization of quality, access, cost and equity.

As discussed in Chapter 5.2, public performance reporting also acts as an indirect economic incentive. The improved reputation that results from credible public performance reporting is a capital asset. Reputation has psychological value to the individual as well as economic value to the individual provider or organization. A better reputation stimulates patient demand and thus confers a competitive advantage, allowing higher prices and higher net income.

Selective incentives (e.g. incremental payments for immunization or screening tests) might be expected to induce a stronger response in the targeted behaviour than a general incentive of equivalent size. For example, capitation payment encourages general efforts in health maintenance and promotion but direct fee-for-service increments for particular preventive and health promotion activities are more likely to lead to increases in those activities.

In economic theory, non-financial incentives are less efficient reinforcers than financial incentives. Whereas the dollar value of a financial reward is identical across persons and organizations, the utility of non-financial incentives such as recognition, administrative simplification and IT grants varies by individual and by organizational context. For example, direct transfer of dollars (cash subsidy) leads to greater improvement in the welfare of the recipient than an in-kind subsidy that costs the grantor the same. The recipient receives more advantage from allocating the dollar between different goods and services according to his/her personal preferences than from a dollar's worth of one particular commodity. Analogously, a producer would rather receive a dollar in subsidy to allocate between different inputs of capital and labour than a subsidy for a dollar's worth of labour. Selective contingent incentives alter the relative price of different activities while general non-contingent incentives provide general rewards or penalties that influence behaviour only indirectly.

Extrinsic and intrinsic incentives

Traditional microeconomic theory does not imply any direct effect of external incentives (such as financial rewards or penalties) on the internal motivation of the individual. The neoclassical model takes consumers' tastes and preferences as given and demonstrates that (irrespective of those subjective values and holding consumer wealth constant) lowering the relative price of any given activity or service will induce the consumer to use more of it. In this model, tastes and personal values such as intrinsic motivation are independent of market conditions and relative prices, for example. Financial rewards and penalties alter the relative prices of behaviour but will not directly affect intrinsic incentives for the same behaviours. Thus, financial incentives for quality improvement would not directly reduce (or accentuate) physicians' inherent interest in optimizing the health benefits for patients.

There are at least two reasons to be cautious about such a conclusion. First, relative price changes (incremental rewards or penalties) have income effects on behaviour by increasing or decreasing provider income. In the balance between net income and the intrinsic payoffs of patient health benefit, as net income rises financial rewards will strengthen the intrinsic motivation of providers who favour patient benefit whereas penalties will weaken it. There is no way to know a priori whether financial rewards reinforce or weaken the provider's intrinsic valuation of patient benefit (Congelton 1991; Frey 1997).

Second, cognitive psychology provides strong evidence that extrinsic incentives (financial and non-financial) crowd out intrinsic motivation (Kohn 1999). Deci et al's (1999) meta-analysis of 128 studies found that all forms of reward – whether contingent on engagement in the activity, completion or level of performance – significantly reduced free-choice intrinsic motivation. Positive feedback (without additional reward) led to increased levels of the activity (free-choice behaviour) as well as self-reported interest in the activity. Deci and Ryan (1985) concluded:

> ... by far the most detrimental type of performance-contingent rewards – indeed, the most detrimental type of rewards – is one that is commonly used in applied settings, namely, one in which rewards are administered as a direct function of people's performance. If people do superlatively, they get large rewards,

but if they do not display optimal performance, they get smaller rewards.

These experimental findings do not imply that financial (and non-financial) incentives will fail to direct behaviour towards the target in the short-term. However, they do raise caution regarding potential long-term negative effects on intrinsic motivation, especially if rewards are accompanied by increased monitoring, assessment and peer competition (Deci & Ryan 1985). The cognitive psychology and industrial psychology literature demonstrates the importance of supporting what Amabile and Kramer (2007) call 'inner work life' – enabling people to progress in their work and treating them decently. Deci and Ryan (1985) argue that intrinsic motivation is grounded in psychological needs for autonomy and competence. Incentive structures that conform to these values would be less likely to undermine intrinsic motivation and are particularly salient for self-regulating professions such as medicine.

Behaviour targeted by the incentive

Narrowly circumscribed incentives oriented on a few discrete tasks or performance measures risk encouraging providers to sub-optimize by multi-tasking or treating to the test (Holmstrom & Milgrom 1991). Such incentives also encourage cream-skimming, i.e. selecting patients for whom it is inherently easier to achieve good performance. Eggleston (2005) demonstrates that mixed payment systems of partial capitation and fee for service will improve performance when incentive contracts fail to specify the full range of provider behaviours necessary to achieve optimal patient outcomes by: (i) muting the adverse effects of incomplete pay-for-performance incentives; and (ii) balancing the cost control incentives of capitation with the quality-promoting potential of fee-for-service payment.

Providers may have differing valuations of patient health benefit relative to net income. Jack (2005) has shown that the best balance between greater provider participation and lower cost to the payer is likely to be achieved by offering provider groups an array of payment contracts with varying degrees of supply side cost-sharing (capitation = 100% provider cost-share; fee-for-service reimbursement approximates 0% cost-share). Providers with higher marginal valuation for

patient benefit will select a higher proportion of fee-for-service payment; those who place greater weight on net income will favour capitation. For example, between 1991 and 1999 general practices in the United Kingdom had the option to become fundholders, receiving a budget to pay for non-emergency, hospital-based specialty care. This voluntary contracting regime captures two points along Jack's schema as fundholding general practices accept partial capitation and non-fundholding practices continue with no direct referral incentives (effectively a type of fee for service). Having adjusted for physician self-selection, Dusheiko et al. (2006) found that fundholding was related to lower hospital admission rates, as expected.

Mixed payment models are designed to address two interrelated performance goals – maximum patient health benefit at least cost. To the extent that policy-makers wish to achieve goals of population access to health benefits and equity, tools other than provider incentives are likely to be more effective and efficient.

Finally, the performance measures that underlie incentives inevitably blend structure, process and outcomes of care (Conrad & Christianson 2004; Kuhn 2003; Young & Conrad 2007). Chalkley and Khalil (2005) show that outcomes-based incentives may be superior when patients are not knowledgeable about their own medical conditions and costs of care but do respond to perceived differences in treatment. Similarly, they demonstrate that outcomes-based payment may be superior for not-for-profit providers who are more intrinsically motivated by patient health benefit.

Policy- and decision-makers deciding on the mix of structure, process and outcome to incentivize must balance the cost and gains of achieving various policy goals. The approach is two-fold (Prendergast 1999):

1. Craft incentives that induce providers not only to treat patients cost-effectively but also, in turn, to reveal their superior information about costs and benefits of different preventive, diagnostic and treatment regimens (incentive compatibility).
2. Pay amounts sufficient at the margin to make providers at least as well off under the incentive regime as they were before (participation constraint).

These two conditions can be satisfied best by predominantly incentivizing behaviour (structure and processes of care) under the proxi-

mate control of providers and also including outcomes in the incentive formulae. On this logic, payment formulae weight the structures and processes chosen as behavioural targets positively (according to the present value of their expected benefits net of costs) and negatively (according to the errors in estimating those net benefits). Other things being equal, process and structure measures strongly related to patient health outcomes (i.e. with large and statistically significant estimated dose-response coefficients) would receive more weight, as would outcome measures that a provider can control more directly and cost-effectively. Conversely, for a given dose-response, measures with less estimating precision would be weighted less in the incentive. In practice, this decision rule places substantial demands on the clinical and economic evidence base but the public, providers and policy-makers should demand nothing less.

Magnitude of incentive

The size of an incentive is optimized by balancing two factors. First, the incentive payment must cover a provider's marginal costs for adjusting behaviour in the targeted direction (Avery & Schultz 2007). This will motivate provider response. There is a subsidiary benefit from tailoring the size of the incentive payment to the marginal cost of performance improvement. When incremental returns (revenues minus costs) are equalized approximately across different dimensions of performance (cost control, clinical effectiveness [quality], patient satisfaction) this attenuates providers' tendency to treat to the test or optimize only certain behaviours. Second, payment should not be higher than is necessary to induce provider participation in the incentive programme. This will contain programme costs by minimizing the 'rents' (payments above marginal cost) captured by providers. Of course, this optimal trade-off is easier to state than to achieve.

Certainty of incentive application

The power of incentives is closely tied to their certainty; the signal-to-noise ratio of incentives is diminished by uncertainty regarding their size, behaviours rewarded, achievability and duration. It is especially important to be clear about the expected duration of incentives and the achievability of underlying performance targets. Incentives that are

expected to be short-term and/or implausible will not stimulate behaviour change, even if they are large and broad-gauged.

Frequency and duration

In principle, more frequent incentive payments will be stronger reinforcers. This reflects the heightened salience that accompanies increased frequency. Also, greater frequency connects reward or penalty more proximately to behaviour and raises the present value of the incentive revenue. The useful life of the provider's investment in quality and efficiency improvement lengthens as the expected duration of the incentive increases, thereby enhancing the expected return on those investments. Moreover, as Kohn (1999) has argued, long-term incentives pose a lesser risk to long-term intrinsic motivation.

Base of comparison: relative versus absolute performance measures

Relative performance measures directly reveal comparative information on providers and, if disclosed publicly, potentially heighten competition. Transparent identification of performance differences also accentuates reputational incentives. Comparative performance incentives adjust implicitly for exogenous shocks common to providers in the same area (e.g. changes in input prices, shifts in area socio-demographics).

The aggregate budget for relative performance incentive payments is fixed by policy and therefore is actuarially predictable (Rosenthal & Dudley 2007). Once the eligible pool of providers is fixed and the structure of rewards and/or penalties is determined then the corresponding incentive budget is known with certainty for a given period. For example, consider an eligible panel of 1000 primary care providers participating in an incentive budget which pays $ 2000 to each provider in the top-performing decile and $ 1000 per provider in the second (80th-89th percentile). In this case the incentive budget equals $ 20 000 (2000 X 10) + $ 10 000 (1000 X 10), or $ 30 000. However, this budgetary certainty is accompanied by uncertainty regarding peer performance which is beyond the individual provider's control. Major gains in quality or cost control may still fall short of the incentive threshold if others achieve even better performance.

Whether payments are increased continuously along a gradient of performance improvement or based on exceeding a specific threshold, absolute performance-based incentives offer providers greater control in attaining the reward. Between the two absolute incentive structures, continuously increasing incentive payments create stronger motivation by avoiding the all-or-nothing property of specific thresholds. Continuously increasing incentive payments account for increasing marginal costs for achieving higher levels of performance (Avery & Schultz 2007; Conrad et al. 2006), strengthening their incentive properties in comparison to relative performance schema. The superior incentive power of absolute performance-based rewards and penalties must be weighed against the greater actuarial uncertainty for incentive payers who must predict the distribution of provider performance and consequent level of payout.

Empirical evidence on performance incentives

This chapter examines performance incentives at two levels. The first is between a health plan (e.g. private insurer in the United States; sickness or statutory health funds in Germany or the Netherlands; general practice partnerships in the United Kingdom) and a provider organization (e.g. medical group practice or independent practice association in the United States; primary care team or general practice fundholder in the United Kingdom). The second is between a provider organization and an individual provider in all health systems. Incentives for the former are determined by health plan payment to providers (general incentives of fee for service, case rates, capitation or a hybrid, coupled with selective incentives for quality or efficiency). For the latter, within the provider organizations, individual physician compensation methods and ownership forms determine the incentive structure.

Health plan to provider organization incentives

The core of this chapter is devoted to selective incentives for quality performance but also presents evidence of how capitation, per case and fee-for-service payments affect physician behaviour. These general incentives establish the overall payment framework within which specific incentives are applied. To date, no published research has compared the effects of selective quality incentives within capitation, per

case and fee-for-service payment regimes. Early pay-for-performance incentives have been applied principally in health maintenance organizations (HMOs). They mitigate the problem of attribution by assigning each enrollee to a particular practice organization or individual provider. This subsection concentrates on the main effects on physician behaviour of general health plan payment methods because the pay-for-performance evidence base does not allow the analyst to isolate interaction effects between general payment methods and selective incentives.

The evidence base in this domain is summarized in two major review papers (Chaix-Couturier et al. 2000; Gosden et al. 2000) and Miller and Luft's (1997 & 2002) reviews in the United States. Chaix-Couturier et al. report that fundholding in the United Kingdom has had no impact on specialist referral or hospital admission rates among general practitioners (Coulter & Bradlow 1993) but has produced consistent reductions in drugs per prescription (Bradlow & Coulter 1993; Himmel et al. 1997; Maxwell et al. 1993; Whynes et al. 1995; Wilson et al. 1996). The shift from fee for service to fundholding led to fewer referrals for elective surgery and to private clinics. Relative to fee for service, capitation payment reduced the number of hospital days by up to 80%.

Chaix-Couturier et al. (2000) synthesized the results of several randomized trials of general financial incentives. Among second and third year paediatric residents Hickson et al. (1987) tested the effect of $ 2 per patient visit (fee for service) against a $ 20 per month salary – payment levels calibrated to yield equal expected income per group, based on historical use rates. The fee-for-service group had significantly more visits per patient; saw their own patients more often (increased continuity); and their patients had fewer emergency room visits. Davidson et al. (1992) assessed the effects of fee-for-service versus capitation (prepaid) payment among physicians participating in the Children's Medicaid programme. Each physician was assigned responsibility for a panel of children. The prepaid physicians' patients had fewer primary care visits; fewer visits to non-primary care office-based specialists; and fewer emergency visits. Assessing the effects of payment method on the care of elderly persons receiving Medicaid, Lurie et al. (1994) found significantly fewer physician visits and inpatient stays and marginally better self-reported general health ($p<.06$) and well-being ($p<.07$) in the capitation group.

Gosden et al. (2000) summarized two other studies of general incentives not captured in the Chaix-Couturier et al. (2000) review. Krasnik et al. (1990) conducted a controlled before and after study of general practitioners in Copenhagen whose remuneration was changed from capitation to mixed fee for service and capitation. Compared to control practices continuing on mixed fee for service/capitation payment, those shifting from capitation to mixed fee for service/capitation demonstrated a significant rise in face-to-face consultations per 1000 patients in the initial six months, followed by a decline in the second six months to rates insignificantly different from baseline. Referrals to specialists and hospital admissions declined more for the intervention group – significantly so by the second six-month period. Compared to the controls, telephone consultations increased significantly more for the intervention group in both post-periods, as did the rate of diagnostic and curative services. Hutchison et al. (1996) found no significant change in hospital-utilization rates among patients of primary care physicians changing from fee for service to capitation (with an additional incentive payment for low hospital-utilization rates) compared to physicians continuing on a fee-for-service basis.

Physician organization-based (group-level) selective incentive studies

This section summarizes the findings of the three most recent structured reviews of the literature on the effects of quality incentives on physician behaviour (Frolich et al 2007; Petersen et al. 2006; Rosenthal & Frank 2006). These are augmented by studies published since the period spanned by those reviews and by earlier literature reviews covering a broader scope of performance measures.

Petersen et al's (2006) review is the most comprehensive, highlighting the effects of selective payment incentives on clinical quality and, secondarily, on access. Overall, they report that explicit quality incentives produced statistically significant quality improvement in two of nine studies at the provider organization level (Christensen et al. 2000; Kouides et al. 1998) and a partial effect in five other studies, i.e. some but not all provider behaviours showed significant improvement (Casalino et al. 2003; Clark et al. 1995; McMenamin et al. 2003; Rosenthal et al. 2005; Roski et al. 2003). Kouides et al. (1998) reported the positive effects on immunization rates of a stepped bonus

per influenza immunization; Christensen et al. (2000) showed an increase in cognitive services interventions by pharmacists in response to enhanced fee for service. Two studies found that group bonuses had no statistically significant effect on cancer screening for women aged fifty or more (Hillman et al. 1998); or on paediatric immunization and well-child visit rates (Hillman et al. 1999).

A recent study of the pay-for-performance incentives applied by Partners Community HealthCare in Massachusetts (Levin-Scherz et al. 2006) demonstrated partial effects. Potential for bonus distribution and return of withholds was associated with increased development of medical management programmes and improved diabetes care processes but no significant impact on paediatric asthma measures.

The financial incentive demonstration of largest scope and incentive size is represented by the General Medical Services contract enacted in the United Kingdom in 2004 (Doran et al. 2006). The results of this new Quality and Outcomes Framework are discussed in more detail in Chapter 4.1 (Lester and Roland 2009). On balance, performance incentives were related to a modest increase in the improvement rate of quality of care (Campbell et al. 2007).

Hospital-based selective incentive studies

Four recent studies of hospital quality incentives complement the physician organization-level studies summarized above. Lindenauer et al. (2007) assessed differential changes in adherence to process quality measures for 10 conditions and 4 composite quality scores in 207 hospitals participating voluntarily in public quality reporting plus pay-for-performance financial incentives and in 406 hospitals participating only in the public reporting initiative. Participating hospitals were part of the CMS/Premier Hospital Quality Incentive Demonstration (HQID). Under this national demonstration programme Medicare hospital inpatient case rates would be increased by 1% for hospitals performing at the 80th-89th percentile and 2% for those at or above the 90th percentile. In comparison with the control group, pay-for-performance hospitals improved significantly more on process measures for acute myocardial infarction; heart failure; pneumonia; and a composite of all ten measures. Baseline performance was inversely associated with improvement – in pay-for-performance hospitals, the composite of all ten measures improved by 16.1% in those with the

lowest quintile of baseline performance and 1.9% for those in the highest quintile (P<0.001). After adjustments for differences in baseline performance and other hospital characteristics, pay for performance was associated with improvements ranging from 2.6% to 4.1% over the two-year period.

Glickman et al. (2007) examined a subpopulation of acute myocardial infarction patients (those with non-ST-segment elevation) in hospitals participating in CRUSADE, a voluntary quality improvement initiative of the American College of Cardiology and the American Heart Association. They compared processes of care and outcomes for the 54 CRUSADE hospitals participating in the CMS/Premier HQID with those of the 446 non-participating CRUSADE hospitals (the controls). The authors found no significant differences in overall improvement between the incentive and control hospitals. However, incentive hospitals did achieve (small but statistically significant) greater improvement than controls in two domains of adherence – aspirin at discharge and smoking cessation counselling. In parallel, the researchers assessed eight guideline-based measures not scored in the incentive programme and found no significant difference in improvement between the incentive and control group. The latter evidence is inconsistent with a hypothesis of treating to the test.

Grossbart's (2006) comparative study of participating pay-for-performance and public reporting only hospitals, affiliated with the Catholic Healthcare Partners health system, identified somewhat weaker effects of the HQID programme. Overall quality scores improved 2.6% more in pay-for-performance incentive hospitals than in other participant (control) hospitals. However, differences were seen solely among congestive heart failure patients, with no significant differences for those with acute myocardial infarction or pneumonia.

A fourth study of hospital quality incentives estimated the cost effectiveness of a voluntary incentive programme adopted by Blue Cross Blue Shield of Michigan (Nahra et al. 2006). In years one to three this programme added up to 1.2% to the participating hospital's DRG (case) rate for the organization's degree of adherence to predetermined heart care guidelines (for acute myocardial infarction and congestive heart failure patients). In year 4 it added up to 2%, contingent on the hospital exceeding the median performance of participant hospitals. This incentive blends elements of relative and absolute performance criteria. There was no comparison group of hospitals but the authors

estimated the cost effectiveness of the incentive programme by summing a programme's administrative costs and incentive payments and comparing these to the estimated QALYs gained by the changes in adherence to heart care guidelines. Nahra and colleagues (2006) concluded that improved guideline adherence saved between $ 12 967 and $30 081 in costs per QALY.

Individual physician-based selective incentive studies

Appraising the external incentives applied to the individual physician, Petersen et al. (2006) indicate that five out of six reviewed studies found significant positive or partial effects (Beaulieu & Horrigan 2005; Fairbrother et al. 1999 & 2001; Pourat et al. 2005; Safran et al. 2000). The initial study by Fairbrother et al. (1999) applied a bonus for improvement from baseline plus enhanced fee for service per immunization delivered. The authors concluded that the stepped bonus improved children's up-to-date immunization status but the enhanced fee-for-service incentive showed no significant effect. In a subsequent study, with an increased bonus for up-to-date immunizations, Fairbrother et al. (2001) reported significant positive effects for both the bonus and the enhanced fee for service.

Safran et al. (2000) conducted a cross-sectional survey of physicians in eight network/independent practice association HMOs. They found that physician financial incentives based on patient satisfaction were associated with higher patient ratings on two of the dimensions of care assessed (access to and comprehensiveness of care) but not to other rated dimensions (continuity, integration, clinical interaction, interpersonal treatment, trust). Pourat et al. (2005) conducted a cross-sectional survey of primary care physicians contracting with Medicaid HMOs in eight Californian counties with the highest rates of Chlamydia trachomatis infection and HMO enrolment. Sexually active females were screened for Chlamydia more often by physicians receiving a salary in conjunction with a quality of care incentive than those paid in other ways (capitation plus financial performance, salary plus productivity, salary and financial performance).

Beaulieu and Horrigan (2005) evaluated the impact of an annual bonus for attaining composite scores exceeding a predetermined target (or for achieving 50% improvement) of process and outcomes of medical care for diabetes patients. Physicians participating in the

incentive programme also were provided with a diabetes registry and met in groups to discuss progress in achieving goals for improvement. Physician performance in the incentive group improved significantly over baseline for five of six process measures and two of three outcome measures.

The study did not formally test the difference-in-differences between the incentive and control groups, but the authors note, 'Improved performance in the study group is an order of magnitude greater than the improved performance in the control group' (Beaulieu & Horrigan 2005, p.1327). For example, changes in the percentage of patients with HbA1c levels ≤ 9.5 between the base year of 2001 and the end of the intervention period of 2002 were 13.9% and 1.8% for the intervention and control groups, respectively. Both absolute and percentage improvements in care process were inversely related to baseline performance. The researchers cautioned that the results could not distinguish explicitly between the effect of the financial incentive and the provision of a diabetes registry and group meetings for tracking progress.

In the sole peer-reviewed study of a relative performance incentive for primary care physicians, Young et al. (2007) evaluated the effect of a 5% withhold. A potential return of between 50% and 150% of the withheld contribution was dependent on the provider's ranking on measures of adherence to four process quality measures of caring for patients with diabetes. Except for a single first-year increase in eye examinations there were no significant differences in pre-intervention and post-intervention trends.

Unintended consequences of performance incentives

One salutary feature of the research on provider performance incentives in health care has been the attention paid to potential unintended consequences. This includes providers' sub-optimizing behaviour such as cream-skimming; stinting on care; or directing exclusive attention to measured performance, to the detriment of important but unmeasured dimensions of care (treating to the test).

Petersen et al. (2006) point to four studies indicating the unintended effects of incentives. Shen (2003) uncovered evidence suggestive of cream-skimming in a Medicaid programme for treating substance abuse. The analysis compared the probability of substance

abuse programme clients being classified as 'most severe' by providers participating in performance-based contracting and providers who were not (the controls). They identified a drop of 7% among clients of participating providers and a rise of 2% among the control group. Three other studies (Fairbrother et al. 1999 & 2001; Roski et al. 2003) found that improved documentation in response to the financial incentive, rather than an increase in preventive services per se, was the source of the positive study findings.

Rosenthal and Frank (2006) cite other examples of unintended consequences. The state of Ohio created financial incentives for increased outreach to persons with severe mental illness – basing the extra payment on the number of such people identified by the provider. The researchers (Frank & Gaynor 1994) concluded that there were increases in the census of such persons identified per provider but found no significant increase in actual treatment for these individuals. A variety of other gaming responses have been documented:

- seemingly intentional miscoding of diagnoses, for provider and/or patient economic benefit (Wynia et al. 2000);
- upcoding of discharge diagnoses in order to enhance hospital reimbursement in response to the incentives of the Medicare hospital inpatient prospective payment system (Carter et al. 1990);
- favourable selection of patients and avoidance of high-cost patients under New York State Cardiac Surgery Reporting System, even with risk-adjustment to control for poorer outcomes of high-risk patients (Burack et al. 1999; Moscucci et al. 2005).

Evidence summary

This section has presented extant empirical research on performance incentives, including general payment incentives and selective incentives in the form of pay-for-performance. The paper's theoretical framework will be used briefly to summarize this evidence.

Nature of the incentive (reward versus penalty)

Empirical studies shed little light on whether penalties or rewards evoke a stronger behavioural response. However, available research does confirm that both negative sanctions and positive rewards induce provider responses in the expected direction. Interestingly, Strunk and

Hurley (2004) report that health plans tend to favour positive incentives (carrots) in lieu of penalties (sticks) in their pay-for-performance programmes.

Target entity (group or individual)

The evidence on which level of incentive exerts more powerful effects on performance is ambiguous. In summarizing the existing peer-reviewed literature, Petersen et al. (2006) observe that seven of nine studies of provider group-level incentives showed positive or partial effects on quality; five of six studies of individual-level studies found positive effects on quality. Frolich et al. (2007) indicated that positive effects were demonstrated in one of three group-level randomized trials and five of seven individual-level studies. Private HMOs appear to be mixing their strategies for levels of incentive (Rosenthal & Dudley 2007). Rosenthal et al. (2006) found that 14% of physician pay-for-performance programmes in commercial HMOs solely incentivize individual physician performance; 61% solely incentivize group-level performance; and the remaining 25% blend the two approaches. Where system failure (rather than individual clinician's deficiencies) is the major source of quality problems, group incentives would be expected to dominate those for individuals, as these figures reflect.

Type of incentive

Extant studies demonstrate that behaviour is influenced by general payment system-level incentives (fee for service, per case, capitation), selective pay for performance and indirect incentives of public reporting. Reviews by Miller and Luft (1997 & 2002) confirm that HMOs' system-level capitation incentives produce somewhat lesser use of hospitals and other expensive resources than do indemnity payments based on fee for service. HMO and non-HMO settings deliver roughly comparable quality of care levels but HMO enrollees report inferior experience on many measures of access to care and lower levels of satisfaction with certain domains, including physician-patient interaction (Miller & Luft 2002). The results for capitation payment in Europe and fundholding in the United Kingdom are consistent with studies in the United States (cf. Gosden et al. 2000; Mossialos et al. 2005).

Extrinsic incentives: effects on intrinsic motivation

No peer-reviewed research of pay-for-performance programmes in health care has estimated the direct impact of selective financial incentives on provider altruism in serving patient needs. Certain forms of sub-optimizing behaviour in response to pay for performance are consistent with diminution in intrinsic motivation: 'treating to the test' (Frank & Gaynor 1994) or avoiding high-cost, low-margin patients (Burack et al. 1999; Moscucci et al. 2005; Shen 2003). At best, these illustrations provide weak evidence of extrinsic rewards crowding out internal aspirations for patient benefit – as Rosenthal and Frank (2006) argue, there are no data to suggest that the pre-incentive overall level of treatment benefits minus costs was superior to the post-incentive level. Glickman et al. (2007) also offer an important counter-example to the posited trade-off of intrinsic for extrinsic reward – non-measured domains of clinical quality did not decline even as certain rewarded types of performance improved.

Nature of behaviour subject to incentive

The first generation of pay-for-performance programmes for physicians emphasized process measures (Petersen et al. 2006) but that is changing. By 2006 over 94% of twenty-four early adopters of pay for performance were using outcomes measures, compared to 59% in 2003 (Rosenthal et al. 2007). No peer-reviewed papers made direct comparisons of outcome- and process-based incentives' effects on actual provider behaviour. However, changes in incentive structure (towards more emphasis on outcomes) constitute survivorship evidence in support of blending outcomes and process incentives.

Only one study (Young et al. 2007) has explicitly evaluated the impact of a relative performance incentive for individual physicians but the authors report no significant effect. Existing pay-for-performance programmes for individual physicians and medical groups favour absolute performance thresholds – 70% of the programmes surveyed by Rosenthal and Dudley (2007). The same survey found that 25% favour pay for improvement, so the predominant pattern in physician pay for performance is one of absolute performance criteria rather than rankings.

Prior studies of physician performance incentives (and the programmes themselves) have targeted preventive services and chronic care. The former reflect the predominance of HMOs in the first generation of programmes; the latter capture the major quality improvement and cost challenges in primary care practice. These clinical domains may offer the most easily achievable quality and efficiency gains but current trends manifest a broadening of the scope of incentives to encompass cost-efficiency, IT and patient experience (Rosenthal & Dudley 2007), as well as specialty practice (Rosenthal et al. 2006).

Incentive size

As Frolich et al. (2007) affirm, previous studies have not identified the dose-response relationship between incentives and the medical care processes or outcomes. The diverse nature of the incentives evaluated and the limited range of variation in the magnitude of any one type (e.g. hospital or medical group, process or outcome, chronic or acute condition) precludes the estimation of robust, precise incentive effects. Petersen et al. (2006) postulate that no or small effects of incentives in several studies (Hillman et al. 1998 & 1999; Kouides et al. 1998) are at least partially attributable to the smallness of the incremental payments.

By combining data on the size of pay-for-performance incentive payments with evidence that previously evaluated programmes have led to modest but typically statistically significant performance improvement it is possible to establish a range for the minimum incentive required to achieve gains. For example, Baker and Carter's (2005) survey of national pay-for-performance programmes indicates that the maximum physician performance bonus was 9%. Rosenthal et al's (2007) look-back interviews of early pay-for-performance adopters reveal that the average physician performance bonus in their sample was 2.3% of total payment. This 2%-9% range in incentive size probably represents an array of tipping points for the first stage of modest change in provider behaviour.

Certainty, frequency and duration of incentive

State-of-the-art empirical work on health-care performance incentives cannot yield direct estimates of the impact of uncertainty in weakening provider response to incentives. Also, available evidence does not allow assessment of the incremental effects on performance of

increased frequency or duration of incentive payment. However, some clues emerge from a small sample of diverse studies. Petersen et al. (2006) indicate that end-of-year payments may contribute to lack of awareness and salience of the bonus, as exhibited in the Hillman et al. (1999) analysis of a paediatric immunization and well-child visit incentive programme. Similarly, lack of frequent performance feedback seemed to inhibit performance improvement in the smoking cessation incentive programme evaluated by Roski et al. (2003).

With no studies of incentive duration, it is possible only to speculate on the size of the boost in quality and efficiency that might be achieved by establishing incentives that would be predictable and endure over a timeframe sufficiently long to prompt providers to make sustained investments in improved clinical infrastructure and care processes.

Implications for research and policy in performance incentive design

This chapter has identified several remaining challenges for empirical research. The research community should develop study designs to differentiate more clearly the performance effects of: (i) distinct types of incentives (financial and non-financial); (ii) group- versus individual-level incentive mechanisms; (iii) external rewards and intrinsic motivation; (iv) process versus outcome measures; (v) varying sizes of incentive payment; and (vi) differences in the certainty, frequency and duration of incentives.

It is imperative to perform side-by-side comparisons of incentives, differing along one dimension at a time. A mix of purposive and randomized controlled trials will be necessary to isolate each key dimension. Also, when experimenting with new incentive arrangements, it is critical that policy-makers collaborate with researchers to design proper pilot demonstrations and monitoring and evaluation mechanisms. This specificity will deliver more targeted information for policy-makers, executives and practitioners as they refine future performance incentives.

The empirical evidence reported in this chapter leads to certain general observations for policy-makers and the design of incentive mechanisms. First, pressures for cost containment in all types of health systems necessitate a type of dynamic budget neutrality in any new quality or cost incentives. Over the long run, resources available

for new incentives are likely to be limited to the rate of growth in the population and input prices for medical care. Accordingly, it will not be possible to sustain incremental rewards for high-performing providers without dampening growth in payments to those attaining lower levels of quality and efficiency. Such reductions are less likely to be perceived as explicit penalties but will send a signal that there is a price premium for quality and efficiency. This reasoning also implies that marginal increases in the rewards for absolute performance are more likely to catalyse quality and cost improvement than relative performance-based incentives.

Second, a mix of group- and individual-level incentive structures will produce the best results, especially if both types are vetted carefully with the professionals and organizations concerned. Quality and efficiency problems are traceable to individual as well as systemic and organizational failures and both levels of structure and behaviour must be confronted. Considerations of sample size and attribution must be addressed in fashioning the optimal mix of organization- and individual-level incentives.

This writer considers that two substantial policy benefits can be achieved by tipping the balance in favour of group-level incentives. Firstly, organizational decision-makers are given maximal discretion to distribute incentives to individual providers in a manner that reflects group norms and practice priorities. This reinforces the salience and professional credibility of any incentive payment (or withhold). Secondly, by directing funds to the group the incentive payers facilitate improvements in the quality and efficiency infrastructure that are necessary conditions for performance improvement.

A third policy recommendation is to follow the natural evolution of incentive implementation. Specifically, process measures for performance incentives should be recalibrated periodically to ensure achievability and consistency with the state of the art. These should be combined with outcome measures that encourage providers to attain results.

Risk-adjustment of patient populations will be increasingly important to the technical and political sustainability of outcomes-based incentive payments. General incentives (as in capitation, per case and fee-for-service payment systems) also interact with selective pay-for-performance and performance reporting incentives. In particular, risk-adjusted and outcomes-adjusted capitation payment could sig-

nificantly reinforce provider response to public reporting and pay-for-performance initiatives.

Different dimensions of performance necessitate distinct incentive structures. Preventive services may be incentivized best by mixing increased fee-for-service payments to individual clinicians with multi-year risk- and outcome-adjusted capitation contracts with the organization. Chronic care management is probably facilitated most effectively by quality-adjusted, salaried compensation to individual physicians, blended with team incentives and organizational capitation.

A substantial body of evidence reveals that significant quality and efficiency improvement is more likely to occur in organized practice settings (McGlynn 2007; Mehrotra et al. 2006; Rittenhouse et al. 2004). Consequently, incentive design should experiment with explicit subsidies for IT and implicit inducements for modest increases in practice scale. For example, implicit incentives for larger-scale practices could take the form of per-provider infrastructure grants that do not compensate small practices for their lack of scale economies in adopting and using advanced technology or in re-configuring practice infrastructure to improve quality or efficiency. Pay-for-performance and performance reporting initiatives targeted at the organization can create a much more robust infrastructure and context for performance improvement than individual physician incentives alone.

References

Amabile, TM. Kramer, SJ (2007). 'Inner work life: understanding the subtext of business performance.' *Harvard Business Review*, (1 May): pp.72–83.

Arrow, KJ (1963). 'Uncertainty and the welfare economics of medical care.' *American Economic Review,* 53(5): 941–973.

Avery, G. Schultz, J (2007). 'Regulation, financial incentives, and the production of quality.' *American Journal of Medical Quality,* 22(4): 265–273.

Baker, G. Carter, B (2005). *Provider pay-for-performance incentive programs: 2004 national study results*. San Francisco, CA: Med-Vantage, Inc.

Beaulieu, ND. Horrigan, DR (2005). 'Putting smart money to work for quality improvement.' *Health Services Research*, 40(5 Part 1): 1318–1334.

Bradlow, J. Coulter, A (1993). 'Effect of fundholding and indicative prescribing schemes on general practitioners' prescribing costs.' *British Medical Journal,* 307(6913): 1186–1189.

Burack, JH. Impellizzeri, P. Homel, P. Cunningham, JN Jr (1999). 'Public reporting of surgical mortality: a survey of New York cardiothoracic surgeons.' *Annals of Thoracic Surgery,* 68(4): 1195–1202.

Campbell, S. Reeves, D. Kontopantelis, E. Middleton, E. Sibbald, B. Roland, M (2007). 'Quality of primary care in England with the introduction of pay for performance.' *New England Journal of Medicine*, 357(2): 181–190.

Carter, GM. Newhouse, JP. Relles, DA (1990). 'How much change in the case-mix index is DRG creep?' *Journal of Health Economics*, 9(4): 411–428.

Casalino, L. Gillies, RR. Shortell, SM. Schmittdiel, JA. Bodenheimer, T. Robinson, JC. Rundall, T. Oswald, N. Schauffler, H. Wang, MC (2003). 'External incentives, information technology, and organized processes to improve health care quality for patients with chronic diseases.' *Journal of the American Medical Association*, 289(4): 434–441.

Chaix-Couturier, C. Durand-Zaleski, I. Jolly, D. Durieux, P (2000). 'Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues.' *International Journal for Quality in Health Care,* 12(2): 133–142.

Chalkley, M. Khalil, F (2005). 'Third party purchasing of health services: patient choice and agency.' *Journal of Health Economics*, 24(6): 1132–1153.

Christensen, DB. Neil, N. Fassett, WE. Smith, DH. Holmes, G. Stergachis, A (2000). 'Frequency and characteristics of cognitive services provided in response to a financial incentive.' *Journal of the American Pharmacy Association*, 40(5): 609–617.

Clark, RE. Drake, RE. McHugo, GJ. Ackerson, TH (1995). 'Incentives for community treatment: mental illness management services.' *Medical Care*, 33(7): 729–738.

Congleton, RD (1991). 'The economic role of a work ethic.' *Journal of Economic Behavior and Organization*, 15(3): 365–385.

Conrad, DA. Christianson, JB (2004). 'Penetrating the "black box": financial incentives for enhancing the quality of physician services.' *Medical Care Research and Review*, 61(Special Suppl. 3): 37S– 68S.

Conrad, DA. Sales, A. Liang, SY. Chaudhuri, A. Maynard, C. Pieper, L. Weinstein, L. Gans, D. Piland, N (2002). 'The impact of financial incentives on physician productivity in medical groups.' *Health Services Research,* 37(4): 885–906.

Conrad, DA. Saver, BG. Court, B. Health, S (2006). 'Paying physicians for quality: evidence and themes from the field.' *Joint Commission Journal on Quality and Patient Safety*, 32(8): 443–451.

Coulter, A. Bradlow, J (1993). 'Effect of NHS reforms on general practitioners' referral patterns.' *British Medical Journal*, 306(6875): 433–437.

Davidson, SM. Manheim, LM. Werner, SM. Hohlen, MM. Yudowsky, BK. Fleming, GV (1992). 'Prepayment with office-based physicians in publicly funded programs: results from the children's Medicaid program.' *Pediatrics,* 89(4 Pt 2): 761–767.

Deci, EL. Ryan, RM (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Deci, EL. Koestner, R. Ryan, RM (1999). 'A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation.' *Psychological Bulletin*, 125(6): 627–668.

Doran, T. Fullwood, C. Gravelle, H. Reeves, D. Kontopantelis, E. Hiroeh, U. Roland, M (2006). 'Pay-for-performance programs in family practices in the United Kingdom.' *New England Journal of Medicine,* 355(4): 375–384.

Dusheiko, M. Gravelle, H. Jacobs, R. Smith, P (2006). 'The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment.' *Journal of Health Economics,* 25(3): 449–478.

Eggleston, K (2005). 'Multitasking and mixed systems for provider payment.' *Journal of Health Economics*, 24(1): 211–223.

Fairbrother, G. Hanson, KL. Friedman, S. Kory, PD. Butts, GC (1999). 'The impact of physician bonuses, enhanced fees, and feedback on childhood immunization coverage rates.' *American Journal of Public Health*, 89(2): 171–175.

Fairbrother, G. Siegel, MJ. Friedman, S. Kory, PD. Butts, GC (2001). 'Impact of financial incentives on documented immunization rates in the inner city: results of a randomized controlled trial.' *Ambulatory Pediatrics*, 1(4): 206–212.

Fama, E. Jensen, M (1998). Agency problems and residual claims. In: Jensen, MC. *Foundations of Organizational Strategy*. Cambridge, MA: Harvard University Press: pp.153–174.

Figueras, J. Robinson, R. Jakubowski, E (2005). Purchasing to improve health system performance: drawing the lessons. In: Figueras, J (ed.). *Purchasing to improve health systems performance*. Maidenhead: Open University Press: pp.44–80.

Frank, RG. Gaynor, M (1994). 'Organizational failure and transfers in the public sector: evidence from an experiment in the financing of mental health care.' *Journal of Human Resources*, 29(1): 108–125.

Frey, BS (1997). 'On the relationship between intrinsic and extrinsic work motivation.' *International Journal of Industrial Organization,* 15(4): 427–439.

Frolich, A. Talavera, JA. Broadhead, P. Dudley, RA (2007). 'A behavioral model of clinician responses to incentives to improve quality.' *Health Policy*, 80(1): 179–193.

Gaynor, M. Gertler, P (1995). 'Moral hazard and risk-spreading in partnerships.' *RAND Journal of Economics,* 26(4): 591–613.

Glickman, SW. Ou, FS. DeLong, ER. Roe, MT. Lytle, BL. Mulgund, J. Rumsfeld, JS. Gibler, WB. Ohman, EM. Schulman, KA. Peterson, ED (2007). 'Pay for performance, quality of care, and outcomes in acute myocardial infarction.' *Journal of the American Medical Association,* 297(21): 2373–2380.

Gosden, T. Forland, F. Kristiansen, IS. Sutton, M. Leese, B. Giuffrida, A. Sergison, M. Pederson, L (2000). 'Capitation, salary, fee-for-service and mixed systems of payment: effects on the behavior of primary care physicians (review).' *Cochrane Database of Systematic Reviews (online),* (3): CD002215.

Grossbart, SR (2006). 'What's the return? Assessing the effect of "pay-for-performance" initiatives on the quality of care delivery.' *Medical Care Research and Review,* 63(Special Suppl. 1): 29S–48S.

Hickson, GB. Altemeier, WA. Perrin, JM (1987). 'Physician reimbursement by salary or fee-for-service: effect on physician practice behavior in a randomized prospective study.' *Pediatrics*, 80(3): 344–350.

Hillman, AL. Ripley, K. Goldfarb, N. Nuamah, I. Lusk, E (1998). 'Physician financial incentives and feedback: failure to increase cancer screening in Medicaid managed care.' *American Journal of Public Health*, 88(11): 1699–1701.

Hillman, AL. Ripley, K. Goldfarb, N. Nuamah, I. Weiner, J. Lusk, E (1999). 'The use of physician financial incentives and feedback to improve pediatric preventive care in Medicaid managed care.' *Pediatrics*, 104(4): 931–935.

Himmel, W. Kron, M. Thies-Zajonc, S. Kochen, M (1997). 'Changes in drug prescribing under the Public Health Reform Law - a survey of general practitioners' attitudes in East and West Germany.' *International Journal of Clinical Pharmacology and Therapeutics*, 35(4): 164–169.

Holmstrom, B. Milgrom, P (1991). 'Multitask principal-agent analyses: incentive contracts, asset ownership, and job design.' *Journal of Law, Economics, and Organization,* 7(Special Issue): 24–52.

Hutchison, B. Birch, S. Hurley, J. Lomas, J. Stratford-Devai, F (1996). 'Do physician payment mechanisms affect hospital utilisation?

A study of health services organizations in Ontario.' *Canadian Medical Association Journal*, 154(5): 653–661.

Jack, W (2005). 'Purchasing health care services from providers with unknown altruism.' *Journal of Health Economics*, 24(1): 73–93.

Kahneman, D. Tversky, A (1979). 'Prospect theory: an analysis of decision under risk.' *Econometrica*, 47(2): 263–291.

Kohn, A (1999). *Punished by rewards: the trouble with gold stars, incentive plans, A's, praise, and other bribes*. Boston: Houghton Mifflin Company.

Kouides, RW. Bennett, NM. Lewis, B. Cappuccio, JD. Barker, WH. LaForce, FM (1998). 'Performance-based physician reimbursement and influenza immunization rates in the elderly: the primary-care physicians of Monroe County.' *American Journal of Preventive Medicine*, 14(2): 89–95.

Kralewski, JE. Rich, EC. Feldman, R. Dowd, BE. Bernhardt, T. Johnson, C. Gold, W (2000). 'The effects of medical group practice and payment methods on costs of care.' *Health Services Research,* 35(3): 591–613.

Krasnik, A. Grenewegen, PP. Pedersen, PA. von Scholten, P. Mooney, G. Gottschau, A. Flierman, HA. Damsgaard, MT (1990). 'Changing remuneration systems: effects on activity in general practice.' *British Medical Journal,* 300(6741): 1698–1701.

Kuhn, M (2003). *Quality in primary care: economic approaches to analysing quality-related physician behavior*. London: Office of Health Economics.

Landon, BE. Normand, SL. Blumenthal, D. Daley, J (2003). 'Physician clinical performance assessment: prospects and barriers.' *Journal of the American Medical Association,* 290(9): 1183–2289.

Lester, H. Roland, M (2009). Performance measurement in primary care. In: Smith, PC. Mossialos, E. Papanicolas, I. Leatherman, S (eds.). *Performance management for health system improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press.

Levin-Scherz, J. DeVita, N. Timbie, J (2006). 'Impact of pay-for-performance contracts and network registry on diabetes and asthma HEDIS® measures in an integrated delivery network.' *Medical Care Research and Review*, 63(Special Suppl. 1): 14S–28S.

Lindenauer, PK. Remus, D. Roman, S. Rothberg, MB. Benjamin, EM. Ma, A. Bratzler, DW (2007). 'Public reporting and pay for performance in hospital quality improvement.' *New England Journal of Medicine,* 356(5): 486–496.

Lurie, N. Christianson, J. Finch, M. Moscovice, I (1994). 'The effects of capitation on health and functional status of Medicaid elderly: a randomized trial.' *Annals of Internal Medicine*, 120(6): 506–511.

Maxwell, H. Heaney, D. Howie, JGR. Noble, S (1993). 'General practice fundholding: observations on prescribing patterns and costs using the daily dose method.' *British Medical Journal*, 307(6913): 1190–1194.

McGlynn, EA (2007). 'Intended and unintended consequences: what should we really worry about?' *Medical Care*, 45(1): 3–5.

McMenamin, SB. Schauffler, HH. Shortell, SM. Rundall, TG. Gillies, RR (2003). 'Support for smoking cessation interventions in physician organizations: results from a national study.' *Medical Care*, 41(12): 1396–1406.

Mehrotra, A. Epstein, AM. Rosenthal, MB (2006). 'Do integrated medical groups provider higher quality medical care than individual practice associations?' *Annals of Internal Medicine,* 145(11): 826–833.

Miller, RH. Luft, HS (1997). 'Does managed care lead to better or worse quality of care?' *Health Affairs*, 16(5): 7–25.

Miller, RH. Luft, HS (2002). 'HMO plan performance update: an analysis of the literature, 1997–2001.' *Health Affairs*, 21(4): 63–86.

Moscucci, M. Eagle, KA. Share, D. Smith, D. De Franco, AC. O'Donnell, M. Kline-Rogers, E. Jani, SM. Brown, DL (2005). 'Public reporting and case selection for percutaneous coronary interventions: an analysis from two large multicenter percutaneous coronary intervention databases.' *Journal of American College of Cardiology,* 45(11): 1759–1765.

Mossialos, E. Walley, T. Rudisill, C (2005). 'Provider incentives and prescribing behavior in Europe.' *Expert Reviews of Pharmacoeconomics Outcomes Research*, 5(1): 1–13.

Nahra, TA. Reiter, KL. Hirth, RA. Shermer, JE. Wheeler, JRC (2006). 'Cost-effectiveness of hospital pay-for-performance incentives.' *Medical Care Research and Review,* 63(Special Suppl. 1): 49S–72S.

Petersen, LA. Woodard, LD. Urech, T. Daw, C. Sookanan, S (2006). 'Does pay-for-performance improve the quality of health care?' *Annals of Internal Medicine,* 145(4): 265–272.

Pourat, N. Rice, T. Tai-Seale, M. Bolan, G. Nihalani, J (2005). 'Association between physician compensation methods and delivery of guideline-concordant STD care: is there a link?' *The American Journal of Managed Care,* 11(7): 426–432.

Prendergast, CR (1999). 'The provision of incentives in firms.' *Journal of Economic Literature*, 37(1): 7–63.

Rittenhouse, DR. Grumbach, K. O'Neil, EH. Dower, C. Bindman, A (2004). 'Physician organization and care management in California: from cottage to Kaiser.' *Health Affairs*, 23(6): 51–62.

Roland, M (2004). 'Linking physician pay to quality of care – a major experiment in the United Kingdom.' *New England Journal of Medicine*; 351(14): 1448–1454.

Rosenthal, MB. Dudley, RA (2007). 'Pay-for-performance: will the latest payment trend improve care?' *Journal of the American Medical Association,* 297(7): 740–744.

Rosenthal, MB. Frank, RG (2006). 'What is the empirical basis for paying for quality in health care?' *Medical Care Research and Review*, 63(2): 135–157.

Rosenthal, MB. Frank, RG. Li, Z. Epstein, AM (2005). 'Early experience with pay-for-performance: from concept to practice.' *Journal of the American Medical Association*, 294(14): 1788–1793.

Rosenthal, MB. Landon, BE. Howitt, K. Song, HR. Epstein, AM (2007). 'Climbing up the pay-for-performance learning curve: where are the early adopters now?' *Health Affairs*, 26(6): 1674–1682.

Rosenthal, MB. Landon, BE. Normand, SL. Frank, RG. Epstein, AM (2006). 'Pay for performance in commercial HMOs.' *New England Journal of Medicine*, 355(18): 1895–1902.

Roski, J. Jeddeloh, R. An, L. Lando, H. Hannan, P. Hall, C. Zhu, SH (2003). 'The impact of financial incentives and a patient registry on preventive care quality: increasing provider adherence to evidence-based smoking cessation practice guidelines.' *Preventive Medicine*, 36(3): 291–299.

Safran, DG. Rogers, WH. Tarlov, AR. Inui, T. Taira, DA. Montgomery, JE. Ware, JE. Slavin, CP (2000). 'Organizational and financial characteristics of health plans: are they related to primary care performance?' *Archives of Internal Medicine*, 160(1): 69–76.

Shen, Y (2003). 'Selection incentives in a performance-based contracting system.' *Health Services Research*; 38(2): 535–552.

Strunk, BC. Hurley, RE (2004). *Paying for quality: health plans try carrots instead of sticks*. Center for Studying Health System Change (Issue Brief No. 82: pp.1–7).

Tversky, A. Kahneman, D (1986). 'Rational choice and the framing of decisions.' *Journal of Business, 59*(4): 251– 278.

von Neumann, J. Morgenstern, O (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Whynes, DK. Baines, DL. Tolley, KH (1995). 'GP fundholding and the costs of prescribing.' *Journal of Public Health Medicine*, 17(3): 323–329.

Wilson, RPH. Hatcher, J. Barton, S. Walley, T (1996). 'Influences of practice characteristics on prescribing in fundholding and non-fundholding general practices: an observational study.' *British Medical Journal,* 313(7057): 595–599.

Wynia, MK. Cummins, DS. VanGeest, JB. Wilson, IB (2000). 'Physician manipulation of reimbursement rules for patients: between a rock and a hard place.' *Journal of the American Medical Association*, 283(14): 1858–1865.

Young, GJ. Conrad, DA (2007). 'Practical issues in the design and implementation of pay-for-quality programs.' *Journal of Healthcare Management,* 52(1): 10–18.

Young, GJ. Meterko, M. Beckman, H. Baker, E. White, B. Sautter, KM. Greene, R. Curtin, K. Bokhour, BG. Berlowitz, D. Burgess, JF Jnr. (2007). 'Effects of paying physicians based on their relative performance for quality.' *Journal of General Internal Medicine*, 22(6): 872–876.

5.5 *Performance measurement and professional improvement*

ARNOLD M. EPSTEIN

Introduction

As many of the preceding chapters have established, measurement is clearly the first step in improving quality of care. If performance cannot be measured, you cannot genuinely determine how well you are doing or whether different approaches to health-care delivery are associated with higher or lower quality. However, measurement is only part of the answer. Most health care is provided by individual clinicians practising in a variety of sites and there will be no predictable and systematic progress in improving quality unless these professionals become engaged in collecting and using performance data to effect change. This chapter focuses specifically on these issues, particularly the relationship between various aspects of performance measurement and professional improvement.

Quality assurance, quality improvement and performance measurement

Historically, quality management was the province of individual doctors, their professional organizations and the state; the latter exercising control largely through licensure (Epstein 1996). Institutional quality assurance developed in the latter half of the twentieth century as a result of the increasing scientific basis of clinical care; complexity of technology; congregation of different sorts of providers (e.g. physicians, nurses, nutritionists, pharmacists) in hospitals and group practice settings; and the advent of accreditation.

Initial quality assurance efforts in hospitals focused largely on structure and process indicators. Analyses of insurance claims data employed to identify providers who overused services for different clinical conditions or in particular clinical circumstances were also deemed quality assurance efforts in some instances. Particularly at

the outset, quality assurance often focused on identifying performers providing low-quality care, the so-called bad apples. Traditional quality assurance probably did not lead to large improvements in the quality of care and was not popular with providers. Undoubtedly at least some of this attitude arose because physicians saw the effort to identify sub-par performers as an attack on their professionalism and autonomy.

Quality improvement arose in part as a counterpoint to traditional quality assurance and has become increasingly important in the last two decades. It builds on managerial and statistical approaches first applied on a wide scale in Japan after the Second World War and rests on seven central ideas (Epstein 1996).

1. Failure to provide optimal care often reflects remediable systemic problems rather than misconduct by individual providers who generally work hard to provide high-quality care.
2. It is essential to encourage teamwork and cooperation because groups of providers dispense complex care in hospitals and medical groups.
3. Quality of care is an organization's product and commitment to quality must be evident throughout the organizational structure and in all personnel.
4. Continuing measurement, characterization of variation and identification of innovative approaches can improve quality of care across the entire performance spectrum.
5. It is crucial to involve patients and workers across the delivery system and to empower them to identify more effective approaches to delivering care.
6. Feedback from health-care 'customers' is an essential part of assessing quality of care and the impact of improvement interventions.
7. Improvement can be performed most effectively in cycles that include the design of new approaches, implementation and continued monitoring of system performance.

Within quality improvement, performance measurement is used to monitor performance; feed data to providers for benchmarking (normative and comparative); and identify high performers or best practices that characterize particularly effective approaches to care. Performance measurement is central to both quality assurance and quality improvement. However, while quality improvement involves a

component of monitoring for poor quality, it places less emphasis on it, unlike quality assurance.

Engaging professionals in quality of care improvement efforts: what does and does not work

Numerous approaches have been used to encourage physicians to change their practice patterns to improve quality of care. Eisenberg and Williams (1981) and Eisenberg (1986) published early reviews of these approaches but these have now been superseded by hundreds of studies and scores of reviews. Some of the most important approaches based on, or incorporating, performance measurement for professional improvement are described below.

Education

Education is possibly the most basic approach to behavioural change. While it need not be combined with performance measurement, evidence of low performance has often been the trigger for educational efforts. Moreover, as described below, failure to catalyse important changes in behaviour through education alone has led to the use of additional strategies that sometimes incorporate performance measurement.

A large range of educational interventions have been extensively studied and reviewed. These include passive traditional educational strategies, usually consisting of didactic educational meetings (e.g. conferences, seminars, lectures) or dissemination of printed educational materials (e.g. publications, audiovisual material). Several factors likely affect the impact of educational interventions on physician behaviour, including the source of the information; presentation format; mode of delivery; frequency and timing of intervention; and specific content (Framer et al. 2003).

A number of studies and reviews suggest that generally the passive dissemination of information (through lecture-based presentations or printed educational materials) has, at most, a small effect on physician practice and patient outcomes (Bero et al. 1998; Grimshaw et al. 2001; Oxman et al. 1995). For example, Browner et al. (1994) examined the impact of a continuing medical education (CME) programme focused on the recommendations of the National Cholesterol

Education Program (NCEP) in the United States. They found that a three-hour seminar had no impact on screening for high serum cholesterol or compliance with guidelines. Even when the educational intervention was intensified by follow-up meetings and printed materials, it failed to elicit change in physician practice. In a major review, Grimshaw et al. (2001) summarized the outcomes of forty-one prior reviews of a wide range of interventions and concluded that passive educational approaches are largely ineffective and unlikely to change physicians' practices significantly.

The development and promulgation of clinical practice guidelines by prestigious professional organizations or other sources may be regarded as a variant of the traditional educational approaches described above, albeit with an intervention that is often regional or national. As with other educational strategies, the passive dissemination of clinical guidelines has often been found to have little impact. For example, Lomas et al. (1989) examined how guidelines recommending reduced use of Caesarian section affected use rates in Canada. A third of the hospitals and obstetricians reported changing their practice as a consequence of these guidelines and obstetricians reported reduced rates in women with histories of a previous Caesarean section. However, data on actual practice showed only a slight decrease. Lomas (1991) also reviewed prior studies of passive dissemination of guidelines and found little evidence that this approach induced change in provider behaviour. Grimshaw et al's (2004) more recent review has similar findings.

Passive strategies alone thus appear to have little impact on physician behaviour but educational strategies that employ interactive methods to engage medical providers can be more effective. Admittedly, the implementation of active approaches may require more resources since they are inevitably more expensive and difficult logistically then simply mailing written materials or publicizing educational information. Thomson O'Brien et al. (2001) demonstrated that interactive workshops that utilize small group discussions and practice sessions can result in moderately large changes in clinical practice. Other studies of active educational approaches such as outreach visits or educational sessions by charismatic opinion leaders have also often shown positive outcomes, although effectiveness varies (Grimshaw et al. 2001; Oxman et al. 1995).

Moreover, multifaceted interventions that use several strategies are generally more effective than single interventions (Grimshaw & Russell 1993; Grimshaw et al. 2001). For example, Headrick et al. (1992) compared three approaches for improving physician compliance with clinical guidelines for the NCEP. Physicians were grouped in three categories (i) standard lecture; (ii) standard lecture + reminder of NCEP guidelines; (iii) standard lecture + patient-specific feedback. This study found that the didactic lectures alone did not improve compliance with NCEP guidelines but the latter two groups experienced some improvement. Box 5.5.1 provides additional examples from the literature of studies incorporating active approaches to education in five countries.

Box 5.5.1 Studies of education coupled with outreach

- In Australia, Cockburn et al. (1992) compared three approaches for marketing a smoking cessation intervention kit to 264 general practitioners: (i) personal delivery and presentation by an educational facilitator; (ii) delivery to receptionist by a volunteer courier; (iii) postal delivery. Doctors receiving the first approach were significantly more likely to see the kit; rate the method of delivery as motivating; use one of the intervention components from the kit; report that they found the kit less complicated; and report greater knowledge of how to use the kit.
- In England, Berings et al. (1994) studied 128 primary practitioners and compared the impact of providing: (i) written information about the indications and limitations of benzodiazepines; (ii) both written and oral information from specially trained general practitioners; (iii) no information at all. The number of benzodiazepines prescribed per 100 patient contacts decreased by 24% among physicians who received both oral and written information; 14% among those provided with only written information; and 3% in the control group.
- In Canada, Lomas et al. (1991) evaluated the education of local opinion leaders as well as audit and feedback as methods of encouraging compliance with a guideline for the management of women who had had a previous Caesarean section. The overall Caesarean section rate dropped only in the opinion leader education group.

Box 5.5.1 cont'd

- In Sweden, Diwan et al. (1995) observed a similar effect for prescribing lipid-lowering drugs in primary care. Health centres that offered four group educational sessions, conducted by a pharmacist, on guidelines for managing hyperlipidaemia showed an increase in the number of prescriptions of lipid-lowering drugs per month compared to the control group.
- In the United States, Stross and colleagues showed the effectiveness of medical education programmes at the community hospital level by training and deploying local opinion leaders whom their peers identified as influential and respected clinicians. One programme resulted in a series of significant positive changes in the management of chronic obstructive pulmonary disease (Stross et al. 1983). Another demonstrated substantial improvement in the utilization of diagnostic procedures and management of patients with rheumatoid arthritis (Stross & Bole 1980). More recently, Raisch et al. (1990) showed that one-to-one educational meetings between prescribers and pharmacists improved the prescribing of anti-ulcer agents for outpatients in a health maintenance organization.

The success of active educational strategies, often using outreach or opinion leaders, has not gone unnoticed in the commercial world. The pervasiveness and perceived impact of these approaches is demonstrated by pharmaceutical companies' common use of representatives who visit physicians in their offices and clinical specialists who are hired to present educational sessions for primary care practitioners on newly developed medications.

Audit, profiling and feedback

The variable and sometimes limited effectiveness of education has been partly responsible for widespread efforts to audit physicians, profile their practice and provide feedback on their performance in relation to their peers. The rationale for this approach is the assumption that physicians will be more willing to change their practice if they learn that their behaviour is far below the norm or some recognized high-quality

benchmark. Sometimes the profiling data are used to characterize performance on indicators of clinical quality (e.g. use of beta blockers for the treatment of acute myocardial infarction) but frequently they are also used to measure 'efficiency', or what is often literally risk-adjusted utilization. These measures might include rates of specialty referral for primary care practitioners; use of radiographic testing; or comparative prescription rates for generic and branded medications.

In the United States, numerous national efforts are underway to capture clinical performance data and provide feedback to hospitals and physicians on comparisons between their performance over time and national benchmarks. For example, the Society of Thoracic Surgeons (STS) has been collecting data since 1989 for the STS National Database. Currently this has over 900 active surgeon participants; in some instances the surgeon's hospital serves as a co-participant. Extensive data are collected for each individual patient undergoing adult cardiac surgery, congenital heart surgery or general thoracic surgery, including pre-operative risk factors; history of previous interventions; specifics on the operative procedure; and post-operative complications. Every six months participants receive a case-mix adjusted outcomes report comparing their practice to regional and national benchmarks. The outcomes report provides longitudinal data on outcomes such as mortality and length of stay by procedure and complexity level. Fig. 5.5.1 provides an example of the national data on length of stay provided by the STS.

In addition to these national profiling efforts, many health plans in the United States collect and distribute data on participating individual doctors and medical groups in an attempt to reduce variation and utilization. For example, in a recent national survey of quality management by more than 240 health plans, Landon et al. (2008) examined the collection of data for 7 quality indicators included as part of the HEDIS battery (e.g. screening for breast cancer, control of high blood pressure). Depending on the quality indicator, they found that 50% to 81% of health plans collected quality performance data on individual doctors or medical groups and 38% to 69% of health plans reported these data back to the providers responsible.

The compelling rationale for audit and feedback and its broad use in patient care organizations might imply that it is a highly effective strategy for changing physicians' behaviour. However, early studies in the 1990s indicated that audit and feedback was neither a consistent

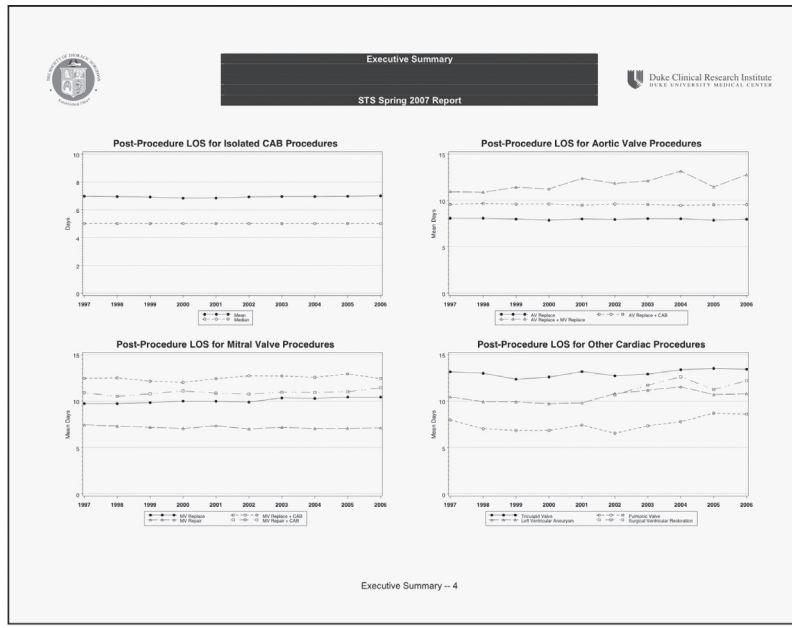


Fig. 5.5.1 Sample of length-of-stay report from STS database

Source: STS database (http://www.sts.org/documents/pdf/ndb/1stHarvestExecutiveSummary_-_2009.pdf)

nor a particularly effective intervention (Axt-Adam et al. 1993; Balas et al. 1996). Reviews of the literature by the Cochrane Collaboration initially affirmed that the effects of audit and feedback varied and it was unfeasible to determine which, if any, features contributed to effectiveness (Jamtvedt et al. 2006). More recently, Jamtvedt et al. (2006) undertook a literature review in which they examined 118 randomly controlled studies to determine the impact of audit and feedback, either alone or in concert with various other interventions such as education, involvement of opinion leaders or outreach visits. This review also concluded that audit and feedback on performance generally has a small to moderate impact. Greater changes occur when there is low baseline adherence to recommended practice and when feedback, with or without educational meetings, is given more intensely (e.g. more frequently). Boxes 5.5.2 and 5.5.3 provide examples from five countries of prior studies of audit and feedback that have addressed different clinical areas or that have been combined with differing interventions.

Box 5.5.2 Studies of audit and feedback by area of health care

Pathology and radiology

In the Netherlands, Buntinx et al. (1993) compared three feedback methods to improve the quality of cervical smears among 179 doctors. Cytologists judged the smears on a three-point scale. Feedback of increasing intensity was provided to: (i) low-intensity group – received written feedback on the technical quality of their sample; (ii) medium-intensity group – received same written feedback plus monthly summaries of their quality performance relative to their peers; and (iii) high-intensity group – received both forms of written feedback plus specific advice concerning their deficiencies. A positive but not statistically significant correlation was observed between improvement in the quality of cervical smears and the increasing intensity of the feedback.

Operative procedures

In the United States, Ferguson et al. (2003) examined the effect of a multi-faceted set of low-intensity interventions to increase the use of beta blocker therapy and internal mammary artery grafting in patients undergoing CABG surgery. Three types of interventions were used: (i) call-to-action by a physician leader; (ii) educational products; and (iii) nationally benchmarked, longitudinal, site-specific feedback. The intervention groups showed modest increases in the use of both process measures, with a significant impact at lower-volume CABG sites.

Prescribing

In Australia, O'Connell et al. (1999) examined the impact of unsolicited written and graphical feedback on the prescribing patterns of over 2000 general practitioners practising in non-urban settings. The test group received mailed, unsolicited, graphical displays of their prescribing rates for two years relative to those of their peers, in addition to educational letters on prescription issues. The authors found no significant change in the prescription patterns of the participants overall or within the subgroups of high and low prescribers.

Box 5.5.3 Studies of different audit and feedback approaches

Audit & feedback with guidelines

In Denmark, Søndergaard et al. (2003) studied the impact of feedback on general practitioners' prescriptions for antibiotics for respiratory tract infections. The control group received clinical guidelines only; the intervention group received guidelines coupled with data on prescription rates versus county averages for various classes of antibiotics. The addition of feedback on prescription patterns failed to change general practitioners' behaviour significantly.

Audit & feedback with education

In Canada, Pimlott et al. (2003) studied how feedback in combination with educational materials affected the rate of physician prescriptions for benzodiazepines in elderly patients. The intervention group received evidence-based educational bulletins and profiling for benzodiazepine prescriptions written for elderly patients. The control group received similar educational materials and profiling for antihypertensive drug prescribing for elderly patients. The authors found that the feedback intervention produced no significant change for either total benzodiazepine prescription rates or for rates of benzodiazepine prescriptions in combination with other psychoactive medications.

Audit & feedback using a multi-faceted approach

In the Netherlands, Verstappen et al. (2003) examined how a multi-faceted approach to audit and feedback impacted on the test ordering performance of primary care physicians. Two test groups focused on different clinical problems (Group A: cardiovascular and abdominal complaints. Group B: chronic pulmonary disease and asthma; general complaints; and degenerative joint complaints). Both groups received mailed feedback benchmarking their test ordering practices against their colleagues. This feedback was followed up with dissemination of national evidence-based guidelines and with regular small group meetings on quality improvement. The study found an improvement in physicians' test ordering

Box 5.5.3 cont'd

practices in both study groups and Group A showed a significant reduction in the number of inappropriate tests ordered.

In the United States, Soumerai et al. (1998) examined how clinician education by local opinion leaders and performance feedback impacted on improving the quality of treatment of acute myocardial infarction. The intervention group received feedback on adherence to treatment guidelines and took part in small and large group educational discussions on treatment guidelines with a local opinion leader. The control group received only mailed feedback on adherence to treatment guidelines. The use of local opinion leaders accelerated the adoption of some beneficial therapies (e.g. aspirin, beta blockers) but had no significant impact on the use of effective but riskier treatments (e.g. thrombolytics for elderly patients).

Accreditation and recertification

Increasingly, quality performance measurement is incorporated into individual and institutional providers' requirements for accreditation and recertification. For the latter, mandated performance measurement is often used as a method for focusing survey processes on substandard or deficient performance areas.

In the United States, the two major accreditors of provider institutions (NCQA and Joint Commission) require the submission of performance data for health plans and for hospitals. NCQA requires health plans to submit both HEDIS and CAHPS. The HEDIS battery includes seventy-one indicators covering eight domains and is described in more detail below. The national oversight committee from NCQA reviews on- and off-site survey team evaluations and performance scores on HEDIS and CAHPS and assigns accreditation ratings in the form of a star system. At present the HEDIS-CAHPS results account for approximately 35% of the overall accreditation points.

Since 2004, the Joint Commission has required hospitals to submit data on three (increased to four in 2008) standardized core measure sets. Each set is a group of indicators covering one of five clinical conditions: (i) acute myocardial infarction; (ii) congestive heart failure; (iii) pneumonia; (iv) surgical infection prevention; and (v) pregnancy and related conditions. The Joint Commission provides a summary

of the reported data using statistical process control techniques for organizational and surveyor use; populates a management tool that compares organizational performance against self-selected cohorts for organizational use; and publishes the data on the Internet (www.qualitycheck.org). The Joint Commission uses performance measurement data to help identify clinical service groups and prioritize focus areas for the on-site survey process. Performance on HEDIS-CAHPS and the core measures reflects overall health plan and hospital performance.

While effective institutional quality management is central to high performance, it would be very difficult for health plans and hospitals to improve without the cooperation of individual providers. Certification of individual health-care providers is gaining attention in multiple countries. One early innovator is The American Board of Medical Specialties (ABMS). In 2002, the ABMS approved a new framework for the maintenance of certification comprising four components: (i) evaluation of clinical performance; (ii) maintenance of an unrestricted licence; (iii) evidence of lifelong learning; and (iv) passing an examination of medical knowledge. The twenty-four specialty boards overseen by the ABMS are required to have recertification programmes that conform to this framework by 2010.

The American Board of Internal Medicine (ABIM) implemented its new programme in 2006. All physicians seeking their ten-year recertification must complete a four-step practice improvement module (PIM): (i) collection of practice data from some combination of medical record audit, patient surveys and a survey about clinical management in their practice; (ii) generation of quality performance measures for review by the physician; (iii) selection of a performance measure to improve, implementation of a strategy to accomplish improvement, and conduct of a rapid cycle test of change involving a small sample of patients over a relatively short period (e.g. several weeks); and (iv) physician's reflections on the impact of the improvement plan and indication of further changes that are intended. The PIMs focus on common issues and concerns such as diabetes, hypertension and preventive cardiology. To date more than 11 000 physicians have completed one of the PIMs and some preliminary data are available about the acceptability of the recertification programme and the quality indicators used in the diabetes PIM (Holmboe et al. 2006; Lipner et al. 2007). However, there is still a lack of information about the approach's success in teaching quality improvement techniques or actually leading to improved care.

Publicly released performance data

The performance measurement efforts described above (perhaps excluding accreditation) are employed largely for internal purposes – to guide quality assurance and quality improvement within healthcare organizations. However, concerted efforts to develop and disseminate publicly information on quality indicators over the last fifteen years have also engendered public, standardized reports on quality of care, commonly known as quality report cards. In the United States, these sorts of data are available much more commonly for hospitals or health plans than for medical groups. Public performance reporting is discussed at length by Shekelle (Chapter 5.2) but this chapter provides brief descriptions of the key reports, targeting United States' hospitals and health plans specifically.

The Hospital Quality Alliance (HQA) is arguably the most extensive current effort to measure hospital quality of care. Developed by a consortium including the CMS, the Joint Commission and the American Hospital Association, since 2003 the HQA has provided regular public reporting for an increasing number (now over twenty) of process indicators of clinical quality for acute myocardial infarction, pneumonia, congestive heart failure and surgical care. Hospitals report these measures on a voluntary basis but the CMS has provided financial incentives for reporting a subset of the measures since 2004. As a result almost all hospitals with sufficient numbers of patients provide the data. The HQA data set has expanded to include risk-adjusted mortality after acute myocardial infarction, congestive heart failure and pneumonia and the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS).

The gold standard for performance measurement of hospitals and individual physicians is perhaps the regular release of statistics on risk-adjusted mortality due to coronary artery bypass graft (CABG) surgery for hospitals and individual surgeons in New York State and Pennsylvania. These data have been made public in periodic reports since the early 1990s. Several other states (e.g. California, Massachusetts, New Jersey) have developed similar systems, although the data from California and New Jersey are at the hospital level only. The CABG reports are notable because researchers have relatively long experience with them and they incorporate extensive efforts to risk adjust the mortality data.

Overseen by NCQA, HEDIS has been the most commonly used report card for health plans for more than fifteen years. The HEDIS battery includes not only information on quality of care but also access to care, enrollees' satisfaction with care and utilization of services. In 2008 HEDIS included indicators covering twenty-three clinical conditions addressing overuse, misuse and underuse of care. HEDIS data released in 2007 included performance results from more than 500 health plans and 80 million HMO, point of service (POS) and preferred provider organization (PPO) enrollees.

Providers' response to report cards

Quality report cards are designed with audiences other than providers in mind – patients who might use them to select providers; large-scale purchasers of health care for contracting or commissioning; and regulators of care who might use them to assure accountability. Each of these audiences may use the data somewhat differently but it is hoped that all of these efforts will result in improved quality performance among health-care professionals.

Evidence suggests that hospitals and health systems (and presumably the doctors and medical groups that populate them) often respond to publicly released data with efforts to improve on measured aspects of care. For example, studies have documented substantial improvement in risk-adjusted mortality after CABG surgery in New York and several other states have initiated public reporting of these data (Hannan et al 1994; National Committee for Quality Assurance 2004). Similarly, NCQA's public release of serial HEDIS data on health plans has been associated with fairly broad improvement in the publicly released indicators. However, success has been variable and some areas (e.g. mental health) have proved intransigent.

Furthermore, even when public performance reports catalyse quality improvement by providers, some critics have raised concerns about unintended responses. For example, Green and Wintfeld (1995) reported data from New York State showing that surgeons began to report higher rates of co-morbidities for their patients after the CABG mortality reporting system was introduced, perhaps leading to a factitious reduction in risk-adjusted mortality over time.

A survey of Pennsylvania cardiologists showed that most respondents thought that the risk adjustment was inadequate and that surgeons and hospitals might manipulate the data to their benefit (Schneider & Epstein 1996). Only 13% of those cardiologists surveyed considered that the reporting system had a moderate or substantial influence on their referral recommendations.

Others have worried that a focus on publicly reported quality indicators will cause physicians to ignore the performance and improvement of other important but unreported aspects of care. Two recent studies have tried to address this concern (Glickman et al. 2007; Landon et al. 2007). They found no evidence of such negative spillovers but the possibility of this kind of skewed emphasis remains.

Finally, there has been substantial concern about potential inequity of care; specifically that physicians or hospitals might limit access to care for patients with greater severity of illness or higher levels of co-morbidity which cannot be addressed fully by the risk adjustment. Studies in the United States have linked better performance on health plan quality indicators with white race and higher socio-economic status (Zaslavsky & Epstein 2005; Zaslavsky et al. 2000) giving rise to concern that patients from racial minorities and lower socio-economic groups also may be at risk of exclusion.

Thus far these concerns about access have been difficult to study or document effectively. When surveyed, 59% of the cardiologists in Pennsylvania reported more difficulty finding a surgeon for severely ill patients needing CABG surgery after adoption of the public reporting system on risk-adjusted CABG mortality; 63% of those surveyed said that they were less willing to operate on such patients (Schneider & Epstein 1996). Omoigui et al. (1996) reported that the number of patients transferred to Cleveland Clinic from New York State increased by more than 30% after the initiation of CABG mortality reporting in New York and that these patients tended to be higher risk than patients transferred from other states. Peterson et al. (1998) found no evidence of restrictions in access to care in New York State when they studied national Medicare data. In fact, the severity of illness of CABG patients in the state increased after the adoption of CABG reporting; and New York State residents who sought CABG surgery in other states had lower co-morbidity than those who received their CABG surgery within the state.

Pay for performance

Public reporting has successfully spawned quality improvement but there are many concerns that the rate of improvement in care is still too low. This has produced increasing interest in tying performance measurements to financial incentives. Financial incentives have been used in medicine for many years. For example, as far back as 1990, general practitioners in England began receiving incremental payments for performing immunizations and Papanicolaou smears (Roland 2004). In the United States, health plans have often provided physicians with small incentives based on patients' satisfaction with care or the use of screening measures such as mammography (Epstein et al. 2004). Incentives have grown and are now being applied to a broader set of quality indicators (including structural measures such as the adoption of IT).

Possibly the best known pay-for-performance programme is that adopted for NHS primary care doctors in 2004. This system provides payment for quality indicators related to clinical care for 10 chronic diseases (including diabetes and asthma); organization of care; and patient experience. The average family practitioner had earned between £ 70 000 and £ 75 000 but average gross income rose by £ 23 000 after the pay-for-performance programme was implemented (Doran et al. 2006).

Pay-for-performance systems have been widely adopted in the private sector in the United States. By 2006, more than half of the health plans covering 80% of plan enrollees had adopted pay for performance for physicians or medical groups; a smaller but substantial number adopted them for hospitals (Rosenthal et al. 2006). In some instances, financial payments have been used to provide incentives indirectly. For example, some employers have incorporated financial incentives in the form of tiering arrangements – patients pay more for providers with lower quality performance or efficiency. Even the federal government has served notice of its interest in a pay-for-performance approach. In the Deficit Reduction Act of 2005, Congress mandated the Secretary of Health and Human Services to develop plans for incorporating performance incentives into the Medicare programme for hospitals by 2009.

Despite the considerable interest in pay for performance, the data on its effectiveness are inconclusive. Petersen et al. (2006) found mixed

results in seventeen studies published between 1980 and 2005, with few strongly positive findings. Four of the studies reviewed showed unintended effects of pay-for-performance programmes (including adverse selection and improved documentation) rather than improved quality of care. Only one study examined cost effectiveness. No studies examined whether improvements in quality persisted over a long period or changes in quality of care as measured by overuse. Similarly, Campbell et al. (2007) found that the quality of care had been improving for diabetes, asthma and congestive heart failure before pay for performance was implemented. The new NHS programme modestly accelerated improvement for diabetes and asthma but not congestive heart failure. Conversely, the same study demonstrated that there was no difference in the rate of improvement between specific clinical indicators associated with financial incentives and unassociated indicators. However, the authors caution that the NHS study was not designed specifically to analyse the difference between indicators with and without incentive attachments and therefore this finding per se cannot be interpreted as proof of the pay-for-performance programme's ineffectiveness.

Two recent studies of a voluntary demonstration programme by CMS in the United States were equally inconclusive. Starting in the last quarter of 2003, hospitals that chose to participate in the Medicare demonstration were eligible for an increase of 2% in their Medicare payments if they reached the top performance decile for one of five clinical conditions: congestive heart failure, acute myocardial infarction, pneumonia, total hip replacement and total knee replacement. Hospitals reaching the second performance decile were eligible for an additional 1% payment; hospitals that failed to exceed the performance levels of the bottom 40% by the third year were penalized. Lindenauer et al. (2007) examined care for acute myocardial infarction, congestive heart failure and pneumonia within this programme. They compared this to care provided by a comparison group of matched hospitals with similar characteristics but no monetary incentive to improve and found improvements averaging 4.1% to 5.2% over two years for those receiving the financial incentive. Glickman et al. (2007) examined acute myocardial infarction using a different comparison group and found no statistical impact from the financial incentives.

In short, review of the literature to date shows clearly the lack of conclusive data on the effectiveness of pay for performance. It seems

likely that multiple factors impact on the success of efforts to spur improvement with financial incentives. These include the nature of the clinical conditions targeted; the size and shape of the incentive programme; and the time lag between initiation of the programme and the measurement of care. All that can be said with confidence is that performance incentives certainly have the potential to work but also the potential to fail.

Quality measurement to encourage professional participation

If performance measurement is to prompt professional improvement, the specific types of indicators used are likely to be as important as the approach through which they are employed. In particular, it seems that physicians are most likely to find indicators acceptable and useful if they serve the functions listed below.

- *Reflect meaningful aspects of clinical practice with strong scientific underpinning.* The most credible indicators are those that reflect important aspects of what physicians perceive that they do; are statistically reliable; and have strong scientific evidence of validity.
- *Assure close risk adjustment of outcome indicators and specify process indicators.* Professionals are intimate with the clinical and social characteristics of patients that lead them to choose different diagnostic and therapeutic approaches. The plaintive refrain, 'my patients are sicker,' accompanies almost every effort in practice profiling. Physicians recognize that outcomes are critically important but the ability to specify process measures more closely to fit a narrow clinical spectrum often makes these more acceptable.
- *Allow exclusions.* Every physician is aware of patients whose medical or social condition made them inappropriate for a particular service, even when they seemed to fit the official clinical profile. The classic complaints concern colorectal screening for patients with dementia, although the problems extend far beyond this. The NHS has addressed this problem by adopting a broad system of exclusions from performance measurement – physicians can exclude patients with atypical clinical situations and for whom performance scoring would be misleading. Proponents of this approach argue that it

has enabled the NHS in England to garner physician support and thereby increase the validity of the performance measurements.

- *Facilitate interpretability.* Process measures are most effective when they indicate clearly what physicians need to do to improve performance. Professionals are likely to mistrust process measures where it is not clear whether higher or lower means better quality of care. Measures such as the proportion of generic medications fall into this category – greater use of generics is often preferable but 100% use is clearly too high. Measures such as these can be confusing and less effective in spurring improvement.
- *Represent services under a provider's control.* Clinicians are most comfortable with quality indicators for which measured performance does not depend greatly on institutional systems or other factors such as patients' compliance. For example, surgeons have complained that risk-adjusted surgical mortality may reflect a hospital's quality of care more than their own individual performance. This may be true, at least for certain procedures. Birkmeyer et al. (2003) have shown that surgical outcomes for some highly technical surgical procedures (e.g. endarterectomy) likely reflect primarily the surgeon's technical skill whereas outcomes for complicated procedures (e.g. pneumonectomy) carried out by operative teams are related more closely to hospital quality.
- *Assure high accuracy.* Health-care providers will strongly favour measures that accurately measure performance. Close specification that yields high reliability; sufficient sample size; and resistance to gaming will all serve to achieve this goal.
- *Minimize cost and burden.* The cost and administrative burden of data collection often falls on the providers who are the subject of performance measurement. Indicators that rely on existing electronic administrative data systems can minimize this burden and thus reduce potential objections.

Policy questions and future challenges for performance measurement and professionals

Performance measurement may be well-advanced but numerous questions and challenges persist.

Should we continue reporting on institutions such as health plans and hospitals or move to performance reports on medical groups and individual doctors?

This question is enormously controversial. In the United States, publicly available performance reports have commonly focused on larger aggregations of providers in hospitals or health plans. This focus reflects easy data availability; the need for adequate sample size; political sensitivity; and concerns about confidentiality. However, there is tremendous impetus to focus on smaller aggregations or even individual clinicians. In England, data are commonly tied to the practice site which generally reflects care by a small number of clinicians. Most patients believe that their individual health-care provider is the person most responsible for their care and data on that provider's practice are the most relevant.

Although systems of care are important determinants of quality and safety, leaders of hospitals and health plans and large practices recognize that they are unlikely to improve quality without the cooperation and changes in the behaviour of individual doctors. Thus far performance measurement seems to reflect acceptable middle ground, with most reports at the individual level remaining confidential. At this point there is no clear consensus about the desirability or practicality of providing a more personal focus.

How can physicians be encouraged to utilize performance measurement and engage more actively in quality improvement?

Part of the answer to this question lies in fostering the use of those quality indicators that are most likely to be acceptable to professionals and employing the strategies that are most likely to engage them. These measures and strategies are discussed at some length above. It would also be helpful to acquaint physicians with performance measurement early in their careers – as a tool to further lifelong professional quality improvement rather than an instrument for inspection and punishment. Better training might also help to foster different attitudes among doctors, encouraging them to recognize their own foibles; the importance of system design in delivering high-quality

care; and the primacy of the needs and health outcomes of their patients. Regulators, accreditors and large-scale purchasers are showing substantial interest in using performance measurement to guide professional improvement. Physicians (and their patients) will benefit if they can be induced to take leadership roles in designing systems to measure and improve the quality of care.

How to create quality indicators to assess specialty care and measure efficiency?

Partly because of the need for sufficient sample size, most quality indicators reflect aspects of care that are very common and under the purview of primary care practitioners. Yet the majority of care, especially expenditure for care, concerns services provided by specialists. For example, in the United States less than 25% of expenditures for office based visits are due to visits to primary care doctors in general practice, family practice or internal medicine (Kurtz 2008). Similarly, most of the process quality indicators employed reflect underuse rather than overuse and exacerbate the growing health-care costs in this country. This trend can be mitigated by introducing more measures of overuse.

Finally, in the last fifteen years the armoury of quality measures has expanded from indicators of appropriate screening and preventive care to a much more comprehensive array of indicators focused on managing chronic disease. These new tools should be used to focus attention on measures that can gauge the performance of specialists and the efficiency of care delivery more specifically.

How to create consortia to better map performance and provide consistent signals?

This is already a particular challenge in countries like the United States, in which physicians contract with multiple payers, and may emerge with increasing use of private insurance in other countries. Several problems may co-exist – significant differences in payers' patient populations can cause scores for the same entity to vary in unexpected ways; no single payer is likely to have enough patients to measure an individual physician's performance reliably without pooling data from other payers; and different specifications for performance indicators

for the same clinical task multiply the administrative burden for those providing the data and may lead to confusing information or false conclusions about performance. These problems have long been recognized but the creation of national (and possibly even international) standards for measures is an ad hoc process that remains a challenge.

When financial incentives are tied to publicly reported data, what are the most appropriate targets (attainment or improvement) and what are the levers that will prompt change most effectively (the magnitude of the incentive or professional ethos)?

Despite considerable experience with pay for performance, many questions remain. Existing pay-for-performance systems show large variations in how they structure incentives, including the magnitude of money at stake and whether targets are tied to attaining certain performance goals or to actual improvement. Rewards for attaining certain performance goals may offer little incentive to improve when providers are already performing well, and may not incentivize very poor performers as they are unlikely to meet goals based on the achievements of the very top performers. Rewards based on relative improvement can be useful – making it possible to reward improvements in very poor performers but disadvantaging those already performing well. These two approaches can be combined in various ways but the resulting complexity and multiplicity of rewards often dilutes the incentive.

There is a need for better understanding of how the magnitude of reward impacts any resulting behavioural change. This is a complicated issue since financial incentives tied to performance provide not only a monetary inducement to improve care but also a signal that draws greater attention to poor performance and the need to improve care. Recent studies have highlighted certain situations in which the signalling function of financial incentives may be particularly important. For example, Rosenthal et al. (2005) and Lindenauer et al. (2007) showed the greatest improvement among providers whose baseline level of performance was so low that they were unlikely to reach the payment target. In these situations the authors concluded that the financial incentives may well have heightened attention to clinical performance and (because of professional ethos) elicited a response from even very

poor performers, particularly in settings where initially low levels of performance facilitated quality gains. Understanding these issues continues to be critically important in programme design and for setting incentives of appropriate magnitude.

Conclusions

In concluding, it seems appropriate to emphasize that performance measurement has become part of everyday life for many practising physicians and already is indispensable in monitoring the quality of care and constructing effective quality improvement efforts. The reality is that none of the methodologies used to date – whether involving confidential profiling; public reporting with aggressive use of incentives; or any other variation – has proven clearly and consistently superior for promoting high quality of care.

Performance measurement is already ubiquitous but many questions and nuances require further exploration in order to increase its usefulness and relevance. Increasing use of IT in health care is likely to make efforts to measure performance even more widespread. The ultimate utility of these efforts will depend on answering the questions and addressing the challenges identified in this chapter.

References

Axt-Adam, P. van der Wouden, JC. van der Does, E (1993). 'Influencing behavior of physicians ordering laboratory tests: a literature study.' *Medical Care*, 31(9): 784–794.

Balas, EA. Boren, SA. Brown, GD. Ewigman, BG. Mitchell, JA. Perkoff, GT (1996). 'Effect of physician profiling on utilization: meta-analysis of randomized clinical trials.' *Journal of General Internal Medicine*, 11:584–590.

Berings, D. Blondeel, L. Habraken, H (1994). 'The effect of industry-independent information on the prescribing of benzodiazepines in general practice.' *European Journal of Clinical Pharmacology*, 46(6): 501–505.

Bero, L. Grilli, R. Grimshaw, JM. Harvey, E. Oxman, AD. Thomson, MA (1998). 'Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings.' *Journal of the American Medical Association*, 317(7156): 465–468.

Birkmeyer, JD. Stukel, TA. Siewers, AE. Goodney, PP. Wennberg, DE. Lucas, FL (2003). 'Surgeon volume and operative mortality in the United States.' *New England Journal of Medicine*, 349(22): 2117–2127.

Browner, WS. Baron, RB. Solkowitz, S. Adler, LJ. Gullion, DS (1994). 'Physician management of hypercholesterolemia. A randomized trial of continuing medical education.' *Western Journal of Medicine*, 161(6): 572–578.

Buntinx, F. Knottnerus, JA. Crebolder, HF. Seegers, T. Essed, GC. Schouten, H (1993). 'Does feedback improve the quality of cervical smears? A randomized controlled trial.' *British Journal of General Practice*, 43(370): 194–198.

Campbell, S. Reeves, D. Kontopantelis, E. Middleton, E. Sibbald, B. Roland, M (2007). 'Quality of primary care in England with the introduction of pay for performance.' *New England Journal of Medicine*, 357(2): 181–190.

Cockburn, J. Ruth, D. Silagy, C. Dobbin, M. Reid, Y. Scollo, M. Naccarella, L (1992). 'Randomised trial of three approaches for marketing smoking cessation programmes to Australian general practitioners.' *British Medical Journal*, 304(6828): 691–694.

Diwan, VK. Wahlström, R. Tomson, G. Beermann, B. Sterky, G. Eriksson, B (1995). 'Effects of "group detailing" on the prescribing of lipid-lowering drugs: a randomized controlled trial in Swedish primary care.' *Journal of Clinical Epidemiology*, 48(5): 705–711.

Doran, T. Fullwood, E. Gravelle, H. Reeves, D. Kontopantelis, E. Hiroeh, U. Roland, M (2006). 'Pay for performance programs in family practices in the United Kingdom.' *New England Journal of Medicine*, 355(4): 375–384.

Eisenberg, JM (1986). *Doctors' decisions and the cost of medical care: the reason for doctors' practice patterns and ways to change them*. Ann Arbor: Health Administration Press.

Eisenberg, JM. Williams, SV (1981). 'Cost containment and changing physicians' practice behavior. Can the fox learn to guard the chicken coop?' *Journal of the American Medical Association*, 246(19): 2195–2201.

Epstein, AM (1996). 'The role of quality measurement in a competitive marketplace.' *Baxter Health Policy Review*, 2: 207–234.

Epstein, AM. Lee, TH. Hamel, MB (2004). 'Paying physicians for high-quality care.' *New England Journal of Medicine*, 350(4): 406–410.

Ferguson, TB Jr. Peterson, ED. Coombs, LP. Eiken, MI. Carey, ML. Grover, FL. DeLong, ER (2003). 'Use of continuous quality improvement to increase use of process measures in patients undergoing coronary artery bypass graft surgery: a randomized controlled trial.' *Journal of the American Medical Association*, 290(1): 49–56.

Framer, AP. Légaré, F. McAuley, LM. Thomas, R. Harvey, EL. McGowan, J. Grimshaw, JM. Wolf, FM (2003). 'Printed educational material: effects on professional practice and health care outcomes (protocol).' *Cochrane Database of Systematic Reviews,* (3): CD004398.

Glickman, SW. Ou, FS. DeLong, ER. Roe, MT. Lytle, BL. Mulgund, J. Rumsfeld, JS. Gibler, WB. Ohman, EM. Schulman, KA. Peterson, ED (2007). 'Pay for performance, quality of care, and outcomes in acute myocardial infarction.' *Journal of the American Medical Association*, 297(21): 2373–2380.

Green, J. Wintfeld, N (1995). 'Report cards on cardiac surgeons. Assessing New York State's approach.' *New England Journal of Medicine*, 332(18): 1229–1232.

Grimshaw, JM. Russell, IT (1993). 'Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations.' *Lancet*, 342(8883): 1317–1322.

Grimshaw, JM. Shirran, L. Thomas, R. Mowatt, G. Fraser, C. Bero, L. Grilli, R. Harvey, E. Oxman, A. O'Brien, MA (2001). 'Changing provider behavior: an overview of systematic reviews of interventions.' *Medical Care*, 39(8 Suppl. 2): 2–45.

Grimshaw, JM. Thomas, RE. MacLennan, G. Fraser, C. Ramsay, CR. Vale, L. Whitty, P. Eccles, MP. Matowe, L. Shirran, L. Wensing, M. Dijkstra, R. Donaldson, C (2004). 'Effectiveness and efficiency of guideline dissemination and implementation strategies.' *Health Technology Assessment*, 8(6): iii–iv, 1–72.

Hannan, EL. Kilburn, H Jr. Racz, M. Shields, E. Chassin, MR (1994). 'Improving the outcomes of coronary artery bypass surgery in New York State.' *Journal of the American Medical Association,* 271(10): 761–766.

Headrick, LA. Speroff, T. Pelecanos, HI. Cebul, RD (1992). 'Efforts to improve compliance with the National Cholesterol Education Program guidelines. Results of a randomized controlled trial.' *Archives of Internal Medicine*, 152(12): 2490–2496.

Holmboe, ES. Meehan, TP. Lynn, L. Doyle, P. Sherwin, T. Duffy, FD (2006). 'Promoting physicians' self-assessment and quality improvement: the ABIM diabetes practice improvement module.' *Journal of Continuing Education in the Health Professions,* 26(2): 109–119.

Jamtvedt, G. Young, JM. Kristoffersen, DT. O'Brien, MA. Oxman, AD (2006). 'Audit and feedback: effects on professional practice and health care outcomes.' *Cochrane Database of Systematic Reviews*, (2): CD000259.

Kurz, R (2008). '7 interesting statistics and facts about office-based physician visits.' *Becker's ASC Review* (http://www.beckersasc.com/news-

analysis-asc/business-financial-benchmarking/7-interesting-statistics-and-facts-about-office-based-physician-visits-by-specialty.html).

Landon, BE. Hicks, LS. O Malley, AJ. Lieu, TA. Keegan, T. McNeil, BJ. Guadagnoli, E (2007). 'Improving the management of chronic disease at community health centers.' *New England Journal of Medicine*, 356(9): 921–934.

Landon, BL. Rosenthal, MB. Norman, SL. Frank, RG. Epstein, AM (2008). 'Quality monitoring and management in commercial health plans.' *American Journal of Managed Health Care*, 14(6): 377–386.

Lindenauer, PK. Remus, D. Roman, S. Rothberg, MB. Benjamin, EM. Ma, A. Bratzler, DW (2007). 'Public reporting and pay for performance in hospital quality improvement.' *New England Journal of Medicine*, 356(5): 486–496.

Lipner, RS. Weng, W. Arnold, GK. Duffy, FD. Lynn, LA. Holmboe, ES (2007). 'A three-part model for measuring diabetes care in physician practice.' *Academic Medicine,* 82(Suppl. 10): 48–52.

Lomas, J (1991). 'Words without action? The production, dissemination, and impact of consensus recommendations.' *Annual Review of Public Health*, 12: 41–65.

Lomas, J. Anderson, GM. Domnick-Pierre, K. Vayda, E. Enkin, MW. Hannah, WJ (1989). 'Do practice guidelines guide practice? The effect of a consensus statement on the practice of physicians.' *New England Journal of Medicine*, 321(19): 1306–1311.

Lomas, J. Enkim, M. Anderson, GM. Hannah, WJ. Vayda, E. Singer, J (1991). 'Opinion leaders vs audit and feedback to implement practice guidelines. Delivery after previous cesarean section.' *Journal of the American Medical Association*, 265(17): 2202–2207.

National Committee for Quality Assurance (2004). 2004 state of health care quality report. Washington, DC: National Committee for Quality Assurance.

O'Connell, DL. Henry, D. Tomlins, R (1999). 'Randomised control trial of effect of feedback on general practitioners' prescribing in Australia.' *British Medical Journal*, 318(7): 507–511.

Omoigui, NA. Miller, DP. Brown, KJ. Annan, K. Cosgrove, D 3rd. Lytle, B. Loop, F. Topol, EJ (1996). 'Outmigration for coronary bypass surgery in an era of public dissemination of clinical outcomes.' *Circulation*, 93(1): 27–33.

Oxman, AD. Thomson, MA. Davis, DA. Haynes, RB (1995). 'No magic bullets: a systematic review of 102 trials of interventions to improve professional practice.' *Canadian Medical Association Journal*, 153(10): 1423–1431.

Peterson, ED. DeLong, ER. Jollis, JG. Muhlbaier, LH. Mark DB (1998). 'The effects of New York's bypass surgery provider profiling on access to care and patient outcomes in the elderly.' *Journal of the American College of Cardiology,* 32(4): 993–999.

Petersen, LA. Woodard, LD. Urech, T. Daw, C. Sookanan, S (2006). 'Does pay-for-performance improve the quality of health care?' *Annals of Internal Medicine*, 145(4): 265–272.

Pimlott, NJ. Hux, JE. Wilson, LM. Kahan, M. Li, C. Rosser, WW (2003). 'Educating physicians to reduce benzodiazepine use by elderly patients: a randomized controlled trial.' *Canadian Medical Association Journal*, 168(7): 835–839.

Raisch, DW. Bootman, JL. Larson, LN. McGhan, WF (1990). 'Improving antiulcer agent prescribing in a health maintenance organization.' *American Journal of Hospital Pharmacy*, 47(8): 1766–1773.

Roland, M (2004). 'Linking physicians' pay to the quality of care – a major experiment in the United Kingdom.' *New England Journal of Medicine*, 351(14): 1448–1454.

Rosenthal, MB. Frank, RG. Li, Z. Epstein, AM (2005). 'Early experience with pay-for-performance: from concept to practice.' *Journal of the American Medical Association*, 294(14): 1788–1793.

Rosenthal, MB. Landon, BE. Normand, SL. Frank, RG. Epstein, AM (2006). 'Pay-for-performance in commercial HMOs.' *New England Journal of Medicine*, 355(18): 1895–1902.

Schneider, EC. Epstein, AM (1996). 'Influence of cardiac-surgery performance reports on referral practices and access to care. A survey of cardiovascular specialists.' *New England Journal of Medicine*, 335(4): 251–256.

Søndergaard, J. Andersen, M. Støvring, H. Kragstrup, J (2003). 'Mailed prescriber feedback in addition to a clinical guideline has no impact: a randomized, controlled trial.' *Scandinavian Journal of Primary Health Care*, 21(1): 47–51.

Soumerai, SB. McLaughlin, TJ. Gurwitz, JH. Guadagnoli, E. Hauptman, PJ. Borbas, C. Morris, N. McLaughlin, B. Gao, X. Willison, DJ. Asinger, R. Gobel, F (1998). 'Effect of local medical opinion leaders on quality of care for acute myocardial infarction: a randomized controlled trial.' *Journal of the American Medical Association*, 279(17): 1358–1363.

Stross, JK. Bole, GG (1980). 'Evaluation of a continuing education program in rheumatoid arthritis.' *Arthritis and Rheumatism*, 23(7): 846–849.

Stross, JK. Hiss, RG. Watts, CM. Davis, WK. Macdonald, R (1983). 'Continuing education in pulmonary disease for primary-care physicians.' *American Review of Respiratory Disease*, 127(6): 739–746.

Thomson O Brien, MA. Freemantle, N. Oxman, AD. Wolf, F. Davis, DA. Herrin, J (2001). 'Continuing education meetings and workshops: effects on professional practice and health care outcomes.' *Cochrane Database of Systematic Reviews*, 1: CD003030.

Verstappen, WH. van der Weijden, T. Sijbrandij, J. Smeele, I. Hermsen, J. Grimshaw, J. Grol, RP (2003). 'Effect of a practice-based strategy on test ordering performance of primary care physicians: a randomized trial.' *Journal of the American Medical Association*, 289(18): 2407–2412.

Zaslavsky, AM. Epstein, AM (2005). 'How patients' sociodemographic characteristics affect comparisons of competing health plans in California on HEDIS quality measures.' *International Journal for Quality in Health Care*, 17(1): 67–74.

Zaslavsky, AM. Hochheimer, JN. Schneider, EC. Cleary, PD. Seidman, JJ. McGlynn, EA. Thompson, JW. Sennett, C. Epstein, AM (2000). 'Impact of sociodemographic case mix on the HEDIS measures of health plan quality.' *Medical Care*; 38(10): 981–992.

5.6 *International health system comparisons: from measurement challenge to management tool*

JEREMY VEILLARD, SANDRA GARCIA-ARMESTO, SOWMYA KADANDALE, NIEK KLAZINGA

Introduction

International comparisons of health system performance provided by multilateral organizations such as WHO and the OECD generate much interest. The provision of comparative data presents vast methodological challenges but offers considerable potential for cross-country learning. Policy-makers are looking for examples, benchmarks and solutions to address the pressures imposed by the epidemiological, economic, societal and technological demands on all European health-care systems.

The use of international performance indicators to assess national economies and public domains such as education, transport and environment has paved the way for their acceptance in the health-care field. Dating back to the 1930s (e.g. Mountin & Perrott 1947), studies on health insurance programmes in western Europe show that international comparisons of health systems were used as a means to guide policy processes (Nolte et al. 2006). Several decades ago, such international assessments focused mainly on structural characteristics (e.g. numbers of physicians, nurses, hospitals) and a few specific outcome parameters (e.g. perinatal mortality, under-five mortality, maternal death, incidence and prevalence of infectious diseases, average life expectancy at birth). In the European region these parameters were complemented by the work on avoidable deaths (Rutstein et al.1976) and release of the first atlas of avoidable deaths in the European Union (Holland 1988 & 1990), thus introducing attempts to assess the contribution of health care to the overall health of populations. Coupled with data on health expenditures (OECD 2001; World Bank 1993),

these produced the first picture on the performance of national health systems in relation to the resources used.

The publication of WHO's *The world health report 2000* and the OECD's *Health at a Glance 2001* received (and continues to receive) much attention. *The world health report 2000* was based on a generic conceptual performance framework and ranked Member States in a league table. Despite many criticisms (see Box 5.6.1), the report placed international health system performance on the political agenda; raised awareness about performance issues; and resulted in many initiatives to improve the perceived health situation in different countries. The latest version of *Health at a Glance* (OECD 2007) contains a comprehensive array of performance indicators without attempting to group the findings in league tables. This has elicited a more nuanced reaction from participating countries. The OECD experience underscores the fact that comparative data help primarily by raising questions about the performance of health-care systems rather than explaining why one country performs better than another.

Box 5.6.1 Debates around *The world health report 2000*

The world health report 2000 was subject to a great deal of controversy. The following points summarize the key controversies pertaining to its political, technical and methodological aspects (McKee 2001):

- Underlying political philosophy – in political and ideological debates the report was accused of being too medical-model based and criticized for its failure to consider the importance of primary health-care systems.
- Face validity – experts questioned the actual rankings of certain countries. For example, the United States ranks higher than Denmark in the responsiveness measure despite the latter having a system of universal health-care coverage.
- Coherence of performance measures – the report was criticized for focusing mainly on health-care systems (instead of considering broader social and educational factors) and not accounting for the lag between health interventions and their measurable impact.
- Data availability – the use of estimates rather than actual data was one of the greatest areas of contention.

Box 5.6.1 cont'd

- Health levels and distribution – critics questioned the use of specific measures such as disability-adjusted life expectancy and equality measures.
- Responsiveness levels and distribution – the use of limited key informants for assessing the responsiveness of health systems and failure to consider the political contexts that could impact this measure was another major area of contention.
- Fairness of financing – critics disputed the definitions and methods used to assess the fairness of financing measures.
- Estimating performance – several debates questioned the 'achievement of performance in health system' concept used in the report.
- Composite index – the use of a composite index (especially the weighting methods used in the report) to measure health systems was heavily questioned.
- Use of evidence – many criticized the report for using a narrow evidence base.

Despite these debates, *The world health report 2000* fostered the importance of health systems. Its publication emphasized the need for health stewardship within national governments and played a significant role in raising the profile of accountability for health on political agendas. Following the release of the report, numerous countries (e.g. Kyrgyzstan) asked WHO for technical support to revise their national health system policies and strategies. Furthermore, it created an impetus for further cross-national discussions around the importance of developing comparable data standards that can be utilized for strengthening health system performance in countries.

This chapter discusses some of the main issues involved in international health system comparisons. The first two sections examine the rationale (why) and the scope (what) of cross-national health system performance assessments, emphasizing the various functions of comparisons (accountability, strategy development, learning) and the scope of such efforts (whole systems, specific services, specific diseases, subnational approaches). Using the OECD's HCQI project as an example, the third section deals with outstanding methodological issues and

challenges (how) such as population variations, data standardization problems, differences in coding practices and definitional issues that arise during international comparisons. The final section addresses the question of how countries can move from measurement to management by illustrating new initiatives that ensure that cross-system data comparisons become an integral part of health system performance management and decision-making processes.

Increased interest in international health system comparisons

Several reasons underlie the increased interest in international health system comparisons. Firstly, policy-makers in resource-scarce environments are increasingly held accountable by the public and the media. International data therefore play a key role in the *accountability* agenda which enables countries to demonstrate that their performance on specific items is equivalent to (or better than) that reported in other countries. Various surveys indicate that accountability can be a generic function of governments towards their citizens but user's negative experiences of health systems can also increase the pressure for governments to seek out best practices and policy lessons from other settings (Schoen et al. 2005). Additionally, the issue of patient responsiveness has recently gained momentum at the European level and could impact on future policy agendas in several countries. Furthermore, patient mobility adds an additional layer of public pressure on governments as borders become more porous in the European region (Legido-Quigley et al. 2008; Rosenmöller et al. 2006).

Secondly, performance information from international comparisons, along with trend data and careful policy analysis, can form the input for national *strategy development* (Hsiao 1992). Following the application of balanced scorecards and strategy maps in the private finance industry (Kaplan & Norton 1992 & 2000), a growing number of countries are in the process of developing frameworks to assess their health systems through national performance reports and strategy development. Examples of such reports are found in the United States (Agency for Healthcare Research and Quality 2008 & 2008a); Ontario, Canada (Veillard et al. 2009); and the Netherlands (Westert & Verkleij 2006). Similarly, the use of balanced scorecards has impacted the establishment of information systems and the management and delivery of health-care services at national and sub-system

levels (Goodspeed 2006; Zelman et al. 2003). International benchmarking data can thus help in formulating the national policy programme. However, it is necessary to use a cautious approach when using comparative data for strategy development purposes since hidden political agendas and selective perception can distort the performance evidence (Klein 1997).

Thirdly, other systems gain opportunities to learn from and emulate the efforts of effective restructuring successes based on performance data from health systems such as the Veterans Health Administration in the United States (Kerr & Fleming 2007). Thus mutual *learning* constitutes the third function of international health system comparisons. As data become more robust it becomes feasible to analyse the factors contributing to better performance – this constitutes an important part of the still limited evidence-based knowledge on health system engineering. The value of sharing similar challenges and experiences is greatly enhanced when governments identify peer groups for comparison. For example, the Nordic Council of Ministers is involved in efforts to compare the quality of care among their countries – Denmark, Finland, Iceland, Norway and Sweden. The results of the study are intended for use in monitoring and evaluating health services while providing a forum for sharing learning experiences amongst participating countries (Wait & Nolte 2005).

In summary, accountability and strategy development are currently the major functions driving governments to engage in international health system comparisons. However, mutual learning is gaining further interest with the increasing scientific robustness of knowledge created through health systems research.

Scope of international health system comparisons

The scope of international health system comparisons varies by country, type of established health information system and availability of resources. The first stage in setting up an international comparison comprises the development or identification of a conceptual framework against which the utility and validity of a set of indicators can be assessed. International organizations have presented conceptual frameworks that aim to describe the underlying constructs and domains and their mutual relations. For example, WHO and the OECD developed such frameworks for health system performance assessment to form

the basis for *The world health report 2000* and a frame for the HCQI project, respectively (Kelley & Hurst 2006) (see Box 5.6.2).

Box 5.6.2 Standardization of performance concepts in international health system comparisons – WHO and OECD conceptual frameworks

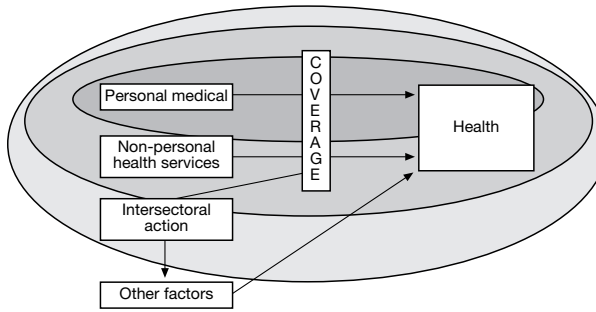
WHO health system performance measurement: WHO chose multidimensional tiers to conceptualize performance, reflecting those considered to be the main goals of a health system – improvement of population health, responsiveness to population expectations and fairness in financial contribution across the population. The main features of this framework are summarized below. Additionally, four main functions were identified (stewardship, financing, service provision, resource generation) in order to provide a relevant policy context for the performance of a health system.

WHO health system performance framework

Components for assessment goals	Average level	Distribution
Health improvement	✓	✓
Responsiveness to expectations	✓	✓
Fairness in financial contribution	–	✓

Source: Murray & Frenk 2000

Boundaries of health systems in the WHO conceptual framework

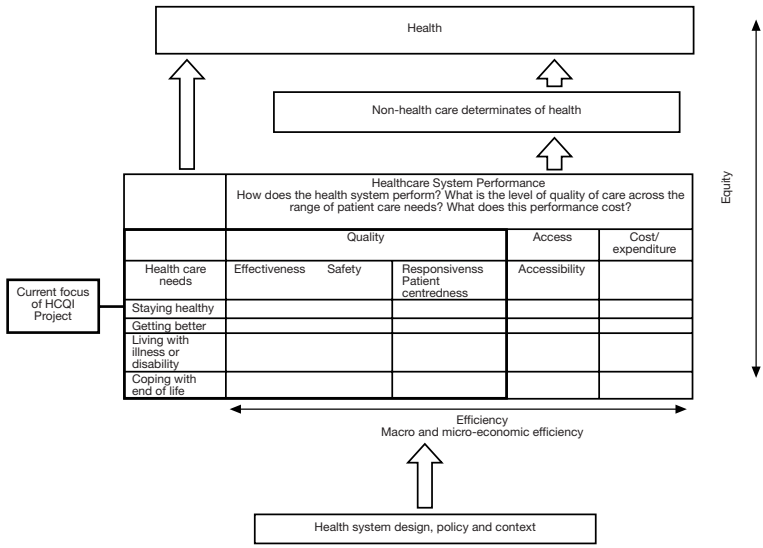


Source: Murray & Evans 2003

Box 5.6.2 cont'd

OECD HCQI conceptual framework: The OECD also adopted a multidimensional approach. The framework below presents a visual summary of the dimensions of health-care performance including: quality, access, cost, efficiency and equity. It also presents a picture of factors related to, but distinct from, health system performance, such as: health system design, policy and context; non-health care determinants of health; and overall levels of health. Finally, it highlights the particular dimensions of quality of care that are the focus of the HCQI project: effectiveness, safety and responsiveness or patient experience.

Conceptual framework for HCQI project



Source: Mattke et al. 2006

The design of a proper set of indicators within such frameworks necessitates the initial, unavoidable task of answering fundamental questions relating to the definition of health system performance, selection of measures and interaction among the individual indicators.

The set cannot be a random list of measures or a simple repository of information and is normally conceived as a system articulating information with a certain purpose – in the case of WHO and OECD, to inform the comparative performance of health systems. There is consensus that indicators selected to compare performance should: (i) be scientifically solid; (ii) be politically relevant; (iii) be available across a sufficient number of countries; and (iv) allow for sustainable and feasible data collection across time (Hurtado et al. 2001; Kelley & Hurst 2006).

The frameworks developed by international organizations encompass structures used in several existing national performance reports and, as Arah et al. (2003) noted, contain many similar dimensions and perspectives. For a classification of the ongoing health system comparisons one can also look at whole system, multilateral, bilateral, disease, sector- or domain-specific approaches. Table 5.6.1 provides a broad categorization of different types of international comparisons of health systems. Some are undertaken on a regular, systematic basis (e.g. OECD HCQI project); others were one-time comparisons (e.g. between United Kingdom's NHS and California's Kaiser Permanente). Although the list is by no means comprehensive, many of these endeavours seek to overcome epidemiological, economic or geopolitical considerations by identifying specific components of the health system and measuring performance on those factors.

As noted earlier, initiatives such as those undertaken by the WHO and OECD assess a broader set of health measures than those studied in traditional comparisons of health systems (e.g. health expenditures among countries; indicators such as life expectancy). Taken a step further, countries and international agencies are increasingly implementing sub-level comparisons, especially at the European Union level. For example, Ben RHM and ISARE are two European Commission funded projects that identified European regions with some common features in their political, socio-demographic and epidemiological development and initiated benchmarking efforts to determine the structural, functional and quality differences of health services within the selected countries. Experiences from these projects show that smaller countries often prefer comparative efforts in which they are evaluated against regions, rather than the entire national health system, of bigger countries (Fédération Nationale des Observatoires Régionaux de la Santé 2007). Furthermore, sub-level comparisons enabled network-

Table 5.6.1 ***General classification of health system comparisons***

Type of initiative	Systems/factors involved	Selected examples
Entire health system	Broad comparisons of overall health systems	• *The world health report 2000*[1] • Commonwealth Fund studies comparing high-performing health systems in the United States, United Kingdom, Canada, Germany, Australia and New Zealand[2]
Multi-lateral	Comparisons between national or sub-national health systems	• Commonwealth Fund study on health system comparisons of six countries that measure various dimensions of health-care systems including quality, access, equity, efficiency and healthy lives[3] • European Commission-funded project: Indicateurs de Santé des Régions Europeénnes (ISARE) covers 283 health regions in 24 European countries[4]
Bilateral	Comparisons between national health systems; national health systems and provincial regional health systems; or national health systems and health-care organizations	• Comparison of health system in Canadian province of Ontario and health system in the Netherlands[5] • Comparison of the United Kingdom's NHS and California's Kaiser Permanente in the United States[6]
Disease-specific	Comparisons of specific health conditions across countries/regions	• Joint WHO/European Commission project: Benchmarking Regional Health Management (Ben RHM) covering 19 regions in 15 European countries and tracking 3 conditions – diabetes, breast cancer and measles[7] • Nordic Council of Ministers' comparisons of specific disease conditions in Denmark, Finland, Iceland, Norway and Sweden[8]

Table 5.6.1 ***cont'd***

Sector-specific	Comparisons of segments of the health-care system e.g. primary care	• Comparison of primary care systems for 18 OECD countries from 1970-1998[9]
Domain-based	Comparisons among components of the health-care system e.g. waiting times, patient experiences	• OECD HCQI project involving 30 countries[10] • Commonwealth Fund study on patient experiences in 7 countries[11]

Sources: [1]WHO 2000; [2, 3]Davis et al. 2007; [4] Fédération Nationale des Observatoires Régionaux de la Santé 2007; [5]Tawfik-Shukor et al. 2007; [6]Feachem et al. 2002; [7]Brand et al. 2007; [8]Wait & Nolte 2005; [9]Macinko et al. 2003; [10]Kelley & Hurst 2006; [11]Schoen et al. 2007.

ing opportunities among health experts and fostered mutual learning experiences (Schröder-Bäck 2007).

A major reform of the health system provides a unique opportunity for countries to undertake comparative studies, allowing related policy and performance changes to be monitored. In 2006, following such a restructuring, the Netherlands initiated a comparative study of their health sector and that of Ontario, Canada, which had undergone reforms during a similar time period. Both Ontario and the Netherlands invested in the development of reliable health system performance assessment frameworks. The study mapped various dimensions of these and compared each of the systems. Conceptual and contextual problems prevent the two systems from being completely comparable but they still provide a starting point for such benchmarking efforts and highlight the range of issues involved in international comparisons (Tawfik-Shukor et al. 2007).

Some researchers have attempted to overcome the larger methodological barriers of cross-country assessments by examining specific components of health systems. For example, a controversial study by Feachem et al. (2002) compared performance factors such as access and responsiveness in the British NHS to the California branch of Kaiser Permanente in the United States. The authors concluded that Kaiser Permanente performed better and had a better integrated and managed system than the NHS, despite similar costs. The study was

heavily criticized for flaws in both its methodology and its assumptions (Himmelstein & Woolhandler 2002) and illustrates that, while individual components of health systems can be compared, it is imperative that such exercises are approached with caution.

This discussion of the various comparative projects is far from complete but illustrates the type of work currently being implemented. In addition, it should be mentioned that major developments are underway to increase the potential of international comparisons in health care at the level of both international research and cross-system databases. At the research level, studies in areas such as cancer care, cardiovascular diseases and diabetes have largely increased the availability of international comparative data. Research projects funded by the European Commission (e.g. Ben RHM, ISARE) are good examples of this type of work currently being implemented. The field of health systems analysis has also expanded and various targeted research groups have been established over the past decade.

Apart from these research processes, expert working groups in international organizations are leading efforts to increase data comparability among countries. Along with WHO's work on the classification of diseases (ICD-9, ICD-10, ICD-11) (WHO 2007) and the OECD's focus on comparing national health accounts and health financing data (OECD System of Health Accounts), there is active collaboration among WHO, OECD and the European Union (Eurostat 2008) to improve the comparability of national data systems.

By contrast, several transition countries in the European region are still establishing their health information systems and therefore comparative studies occur on a limited basis. However, as a first step, a number of countries are involved in the Health Metrics Network (http://www.who.int/healthmetrics) which is hosted by WHO and enables them to overcome problems of data availability and improve the quality and reliability of their information systems. Although some transition countries lack optimal quality control measures, many are increasing investments in efforts to align their health systems with international standards. For example, WHO recently led initiatives by which Armenia and Kyrgyzstan developed performance assessment frameworks to aid them in strengthening their health sectors. In the long run such endeavours will lead to benchmarking among comparable countries in the WHO European Region and highlight areas for improvement in health system performance.

As seen in this section, international health system performance comparisons have a broad scope. Such assessments depend largely on project aims, policy opportunities and the availability of resources and data. Each type of comparison – from multilateral to domain specific – serves an important function in drawing attention to a particular health system and possible ways to strengthen its performance.

Methodological issues in conducting international health system comparisons: lessons from the OECD experience

Initiatives to build relevant and meaningful indicators across different countries face numerous challenges. This section provides an overview of the operational and methodological issues involved in such efforts. The matters explored follow the experience within the OECD HCQI project but can be generalized to comparative efforts in similar international health systems.

The OECD HCQI project started in 2002 with the objective of developing a set of health-care quality indicators that can be reported reliably and regularly across thirty OECD countries. The purpose was to help raise questions for further investigation into differences in the quality of care across countries. The number of countries involved in the HCQI project has recently expanded to include all European Union Member States, including non-OECD nations, following an agreement between the European Commission's Directorate-General for Health and Consumers and the OECD.

The HCQI project has undergone several phases. The initial list of indicators consisted of eighty-six potential measures in five priority areas of care (patient safety; mental health care; health promotion, prevention and primary care; cardiac care; and diabetes care). However, data availability proved to be a major hurdle.[1] There has been a two-pronged strategy to overcome this barrier: (i) initiate regular data collection of widely available indicators; and (ii) simultaneously work with countries to improve information systems and enhance the comparability of indicators. At the current state of development, the regularly updated set covers health areas outlined in Table 5.6.2 (Garcia-Armesto et al. 2007). In addition, fourteen measures for patient safety and two for mental health care have reached the

[1] For a complete description of the short-list building process, refer to entire issue of Mattke et al. 2006.

Table 5.6.2 ***HCQI project indicators***

Care for acute conditions	
Outcome	**Process**
In-hospital acute myocardial infarction case-fatality rates	Waiting times for surgery after hip fracture, age 65+
In-hospital ischaemic/haemorrhagic stroke case-fatality rates	
Cancer care	
Outcome	**Process**
Survival rate for colorectal cancer	Mammography screening
Survival rate for breast cancer	Cervical cancer screening
Survival rate for cervical cancer	
Care for chronic conditions	
Outcome	**Process**
Hospital admission rate for asthma (age 18+)	Annual retina examination for diabetics
Asthma mortality rates (age 5-39)	
Prevention of communicable diseases	
Outcome	**Process**
Incidence of measles	Vaccination against measles
Incidence of pertussis	Vaccination against pertussis (+ diphtheria + tetanus)
Incidence of Hepatitis B	Vaccination against Hepatitis B
	Vaccination against influenza (age 65+)
Other	
Smoking rates	

last phase of piloting and it is envisioned that they will be included in the regular set for 2009 data collection. The indicator set includes both process and outcome measures since they provide different but complementary insights – information derived from process indicators is easier to translate into specific improvements; outcome indicators may be subject to multifactor causal attribution but are indispensable in aligning performance assessment with health system objectives. The key is to establish a balance between these two types of measures.

Within the HCQI project, indicators are considered ready for international comparisons once the agreed threshold of ten countries can provide data from well-identified and stable databases according to agreed definitions (age group, codes, methods of identification). Indicators are added and deleted in order to ensure that the set remains responsive to changes in data availability or measurement quality. The tension between maintaining a stable set over time and the imperative to convey a concise message to policy-makers should be balanced while making decisions about adding and deleting indicators. Furthermore, there is a trade-off between implementing rigorous methodological approaches and including all countries in the calculations. A balance point is achieved when the methodology is strict enough to provide policy insights but flexible enough to allow participation by the maximum number of countries.

Another compromise is to achieve homogeneous information systems without overburdening the countries that are required to comply with such constraints, especially those bearing the cost of adding new data items to their collection structures. The improvement of national health information systems can be considered a positive side effect of involvement in international performance assessment initiatives but any changes must take account of existing structures.

The OECD HCQI project provides rich empirical experience of dealing with complex methodological barriers. Several key issues that need to be considered when establishing and monitoring cross-country performance indicators are listed below.

1. Specifying indicators using internationally standardized definitions.
2. Controlling for differences in population structures across countries.
3. Adjusting for differences in information systems' ability to track individual patients.
4. Controlling variability of data sources.

5. Identifying nationally representative data.
6. Determining retrospective completeness of the time series.

These are described in the following sub-sections together with suggestions to overcome them.

Specifying indicators using internationally standardized definitions

Standardization constitutes the best way to ensure data comparability across countries since it is applied across all stages of data production, storage and report.

WHO leads the main initiative in this field through the WHO Family of International Classifications (WHO-FIC) programme, comprising three types of systems:

1. International Classification of Diseases (ICD)
2. International Classification of Health Interventions (ICHI)
3. International Classification of Functioning, Disability and Health (ICF)

The ICD is used to classify diseases and other health problems and has become the international standard diagnostic classification for epidemiological and health management purposes, ICD-10 is the latest version (an updated ICD-11 is currently under development). However, countries can find it difficult to update to new versions of ICD as its impact in shaping national information systems involves issues such as staff training, adapting to new definitions and changes to funding schemes. For example, ICD-10 contains 12 640 codes while ICD-9 had only 6969. As a consequence, the use of different versions of ICD across countries is a real issue when attempting to identify indicators for international comparison.

In the absence of an internationally accepted system for reconciling ICD-9 and ICD-10, the HCQI project has opted to develop ad hoc validated crosswalks for the indicators relying on them. The first initiative comprises fourteen patient safety indicators that are currently being tested for adoption in 2009. The International Methodology Consortium for Coded Health Information (IMECCHI) is an expert network that has worked with the HCQI project to develop and validate a manual for the calculation of these measures. Consideration of

both ICD versions and the national adaptations of ICD-10 provides a solid basis for 'translation' and enhancing comparability across countries (Drösler 2008).

There are other outstanding issues concerning the calculation of indicators based on standardized codified databases. For instance, actions to address variation in documentation and coding practices across countries will entail some cultural changes that take time. However, participation in international initiatives has the beneficial effect of drawing attention to practices that might be regarded as adequate at the national level, but become less acceptable when compared to those in similar countries.

The current lack of an international classification system for procedures is another relevant aspect, especially for the specification of process indicators. The ICHI covers a wide range of measures for curative and preventive purposes but is still in its beta trial version and entering extensive field trials before being submitted for endorsement by the governing bodies of WHO (WHO 2007). Despite encouraging progress, it may be several years before ICHI is ready for adoption and therefore the HCQI project currently utilizes ICD-9-CM and ICD-10 to specify procedures.

Endorsed in 2001, the ICF seems promising. However, it is not yet used widely across countries and its specific applicability in defining outcome indicators needs to be explored further.

Controlling for differences in population structures across countries

A number of indicators can be affected by a country's demographic structure. For example, survival or mortality rates are influenced by the age and gender structure of the population. This demographic composition has an impact on the epidemiology of diseases and becomes a confounding factor that assessments need to adjust for. Age and sex standardization facilitates comparisons across countries by controlling for these differences in national populations.

When selecting a reference population it is important to decide whether to use the general population or one that is disease-specific (i.e. has the distribution of patients with the respective disease). As the incidence and prevalence of most diseases increases with age, disease-

specific populations tend to weigh older population segments more heavily. A disease-specific reference population is therefore theoretically superior but is frequently not feasible as it requires the construction of a population for each disease. Many research projects overcome this problem by using general population weights. Another technique reduces distortion by removing the segment of the population that is less affected by the disease, truncating the sample to include only those above a certain age, e.g. forty (Lousbergh et al. 2002).

The HCQI project initially considered the 1980 OECD population structure for age-standardization calculations. This decision is now being revised because: (i) the structure of this population is becoming outdated with the demographic ageing trends in OECD societies; and (ii) the OECD has expanded from twenty-four to thirty countries and therefore the 1980 reference has limited validity. The transition to a 2005 OECD reference population is under assessment. The adoption of a truncated population is also being analysed, especially as countries such as Japan face a higher prevalence of myocardial infarction in the elderly group rather than the typical middle-age range.

There is a trade-off in updating the structure of the reference population and maintaining valid comparable data over time. Other international comparative projects face similar challenges caused by ageing populations and incorporating new member countries, e.g. European Union's development of the European Community Health Indicators Monitoring project (2008) or the European Health Interview Survey (2008). Steps should be taken to ensure that the data remain valid and comparable over time.

Adjusting for differences in information systems' ability to track individual patients

Indicators often take the form of rates in which the denominator is a specific group of patients – this cluster of indicators includes hospital fatality rates among patients with certain diagnoses or rates of specific procedures among chronically ill patients. Two interrelated issues affect the feasibility of these indicators: (i) the need to distinguish between different patients and repeated events affecting the same patient; and (ii) the necessity of detecting a patient's contact at any level of care and across different institutions. However, national

information systems do not have a uniform ability to identify patients and often the only data available are activity records which count each episode of care separately, even if the same patient was involved.

There is a clear need to harmonize calculations across countries to ensure data comparability; Mattke et al. (2006) illustrate the effect of different bases of calculation on thirty-day hospital fatality rates for myocardial infarction and stroke. Currently, the most generally feasible approach is events-based calculations in which it can reasonably be argued that the validity of a specific indicator is not affected. However, a unique patient identifier is the most efficient tool for performing patient-based calculations and the OECD recently began encouraging member countries to establish these across their key health information systems.

Controlling variability of data sources

National information systems comprise a variety of data sources with substantial differences in their structure; the nature of data recorded; and the purpose for which they were conceived. Data systems have been shaped to serve monitoring functions within each country. Often, the purpose of such monitoring is neither performance comparison nor quality measurement but rather to support administrative activities such as budget distribution or system management (see Box 5.6.3 for a summary of the main data sources and their general strengths and weaknesses). This means that a fair assessment of the available sources across countries and their suitability (on an indicator by indicator basis) will be required when building indicators for international comparison. For instance, process indicators such as vaccination or screening rates can be built from data from varying sources across countries but the nature of the available data will vary with the structure of health service provisions in each system.

In some countries, prevention activities are organized in large-scale national programmes with routine databases that can be used for analysis. However, data in other countries are managed by each municipality and therefore registries are fragmented and not always accessible at the national level. In addition, registries for prevention activities often do not cover settings outside the health-care system (e.g. work or school) and private organizations that provide this type of care can vary by country, complicating the retrieval of documented

Box 5.6.3 Sources of information available to assess quality of care across countries

Source	Weaknesses	Strengths
1. Administrative data *Admission/ discharge records* *Minimum set of data* *Insurance-reimbursement* *DRGs accounting* *Prescription*	Limited/no information on processes of care and physiological measures of severity Limited/no information on timing (co-morbidities vs. onset or adverse events) Heterogeneous severity within some ICD codes Accuracy depends on documentation and coding Data are used for other purposes, subject to gaming Variation in how administrative data are collected and used, in particular DRG-based payment versus global budgeting versus service-based payment Time lag may limit usefulness Poor development outside the hospital setting	Data availability improving Coding systems (international classifications of diseases) and practices are improving Large data sets optimize precision Comprehensiveness (all hospitals, all payers) avoids sampling/selection bias Data are used for other purposes and therefore subject to auditing and monitoring
2. National surveys *Health status* *Health services use* *Pharmaceutical consumption*	Self-reported (recall bias, lack of accuracy due to lay approach of those interviewed) Inability to identify and follow up subjects	Population based rather than patient based information, including individuals that health information systems cannot account for Can provide a basis for access and needs assessments

Box 5.6.3 cont'd

3. National registries *Cancer* *Chronic diseases* *Adverse events* *Certain procedures* *Mortality*	When not mandatory, some eventual selection bias may deem them not representative Resource intensive to register the detailed specific features (e.g. adding cancer staging data to the diagnosis in cancer registries) Not always linkable to other sources of information	Precise specific information
4. Medical records	Data retrieval is work intensive and therefore expensive, even with electronic records Difficult to sustain over time	Complete clinical information and good chronology
5. Patients surveys *Satisfaction* *Experience* *Access*	Low degree of standardization in patient survey tools, often even within countries Cultural influences on concepts such as satisfaction, expectations and experience hinder comparability across countries	Most reliable method of assessing system responsiveness and obtaining information about how patients perceive and experience the care provided Leads to improvements in designing trans-cultural assessment tools

activity. In other cases, programmes are non-existent and services are provided on a demand basis. In all these situations, population surveys might be the most valid source of information.

The key question is whether data from so many different data sources (registries and population surveys) are comparable. As part of a methodological refinement, the HCQI project assessed the data comparability of surveys and programme registries for cancer screening indicators. Median rates of mammography and cervical cancer screening for each available year were calculated separately for programme and survey data. Based on surveys compared to registries, the variation over time is remarkable and suggests that both sources of data should be utilized with caution. Furthermore, international health system comparisons should use the source factor to adjust differences in the indicators.

Identifying nationally representative data

Cross-national assessments should reflect country-wide data. This is especially true when using process indicators (e.g. measuring care for chronic diseases) where data are often derived from pilots or ad hoc registries and raises serious concerns about the representativeness of data. Unique patient identifiers could make patients much more traceable within routinely collected information and thereby increase the reliability of data collected.

To ensure data comparability across countries, the HCQI project recently adopted a system of classification of the quality of data. This comprises three levels:

- A – corresponds to national administrative registries, with demonstrated non-selection bias;
- B – accounts for non-national administrative registries with demonstrated non-selection bias;
- C – applies to ad hoc registries (e.g. research and pilots) and any other source not classified elsewhere.

Such a system has the advantage of enabling data collection at different levels of quality and using all available data sources, while preserving the rigour of the analysis. For instance, only data within categories A and B can be utilized but C type data can be collected and efforts made to raise them to the two higher categories.

Determining retrospective completeness of the time series

Almost all international comparative efforts face problems in obtaining uninterrupted, reliable data over a given time period. This limits the validity of trend analysis and affects the ability to interpret related indicators together. The time lag between policy implementation (e.g. breast cancer screening for a target population) and expected outcomes (improvement in breast cancer survival rates) can hardly be accounted for in the absence of time series. Prospective time series rely on regularly updated, sustainable data sources; retrospective completeness could be hindered by problems with (for example) the availability of data that need to be considered during international comparisons.

Comparative projects of health systems similar to those developed and implemented by the OECD have great potential in driving health policy. There can be numerous methodological barriers but the process of identifying and overcoming these pitfalls can lead to valid, reliable conclusions that enable effective health decision-making for overall system improvement.

Turning international health system comparisons into health system performance management

International comparisons of health systems can offer governments a valuable tool to revise their policies, review accountability agreements and reassess resource allocation procedures. However, to strengthen health systems it is necessary to use these comparisons for performance management purposes and, as a first step, to integrate performance data needs into the policy-making process. An example from the Ontario Ministry of Health and Long-Term Care illustrates the systematic use of performance information and its flow through the decision-making cycle (Fig. 5.6.1). The diagram shows that comparative data can be used at different stages of the health ministry's business cycle which, as a continuous improvement process, facilitates the use of strategic performance information for performance improvement purposes.

Similar examples can be found in the United States Veterans Health Administration where performance indicators were used to monitor the effects of health system reforms while driving accountability agreements at sub-system and individual levels (Kerr & Fleming

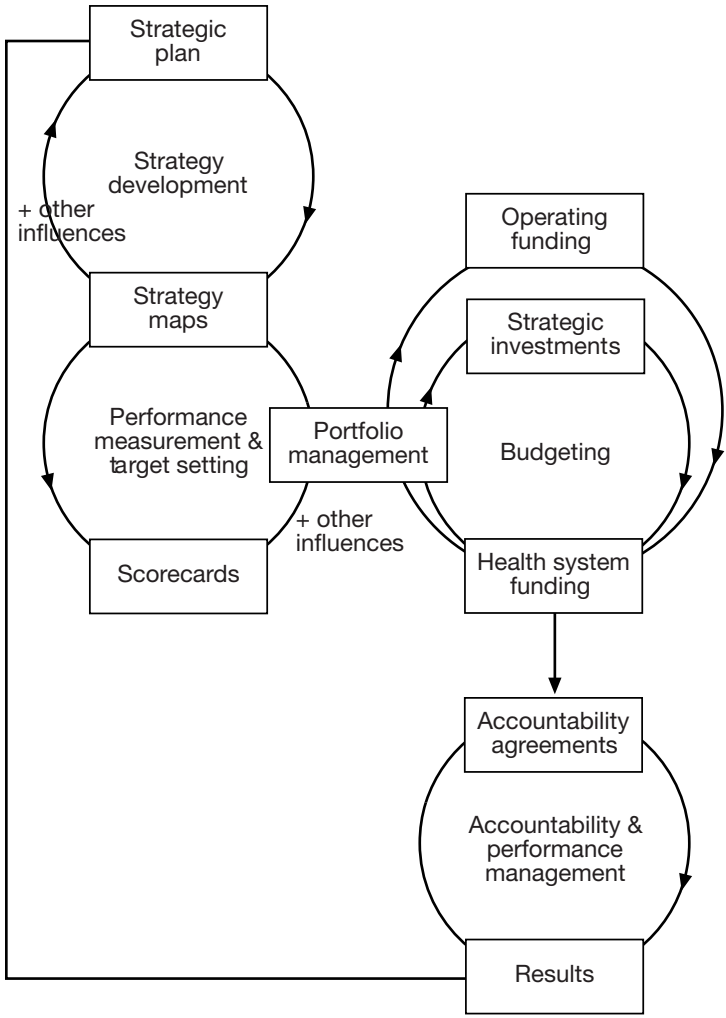


Fig. 5.6.1 Conceptualizing the range of potential impacts of health system performance comparisons on the policy–making process

Source: Veillard et al. 2009

2007). Other successful case studies range from health-care organizations (Kaplan & Norton 2005) to private industry (Kaplan & Norton 2000). In order to guide health policy-makers in the delivery of better results, it is critical to turn strategy-based performance information into performance management systems.

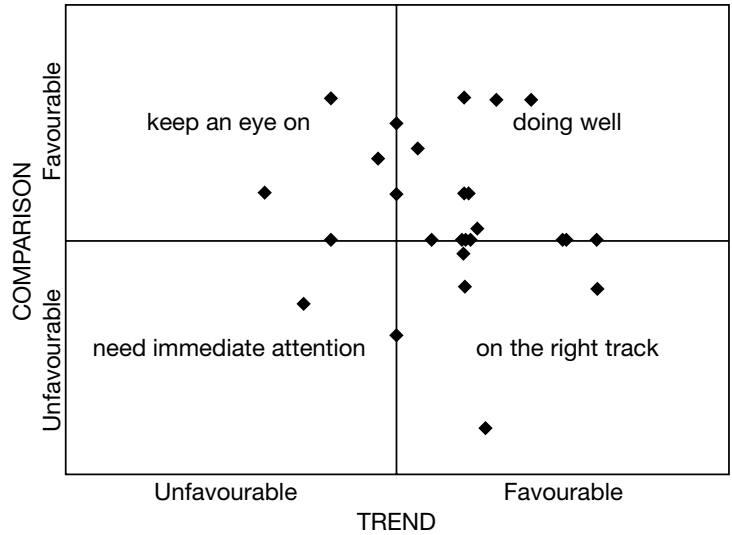


Fig. 5.6.2 Translating benchmarking information to policy-makers. Example from the Ministry of Health and Long-Term Care, Ontario, Canada

Source: Health Results Team for Information Management 2006

Translating performance information for policy-makers

Another crucial aspect of performance management is translating performance information to make it simple and clear to policy-makers (Lavis 2006). For instance, the Ontario Ministry of Health and Long-Term Care represented health system performance measures from two different perspectives: variation in performance over time and against selected benchmarks (or comparators), respectively. These approaches are interesting examples of how to present performance information to health policy-makers in relevant ways. For instance, Fig. 5.6.2 indicates to Ontarian decision-makers whether performance is improving; if it is favourable compared to pre-defined benchmarks (standards, international comparators, provincial comparators); and the policy actions required for different levels of performance. This approach suffered from standardization difficulties but with comparable performance data can be a promising practice for governments wishing to benchmark their health system performance in a concrete fashion.

Funnel plots are another tool for benchmarking performance management and are used increasingly by countries such as the United

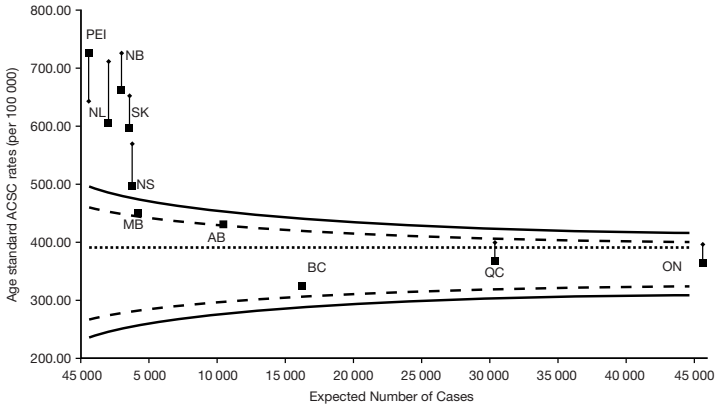


Fig. 5.6.3 Funnel plots for ambulatory care sensitive conditions for different Canadian provinces, 2006 data

Source: Health Results Team for Information Management 2006

Kingdom and Canada (Spiegelhalter 2005). Fig. 5.6.3 shows a set of funnel plots that represent the performance indicator (in this case, rate of ambulatory care conditions) with deviations from the average. A trend component is incorporated by using an arrow to indicate whether performance has improved or declined; the length of the arrow shows the relative magnitude of change over time. The calculation of funnel plots is associated with some statistical problems but they can provide policy-makers with a visual representation of their country's relative performance against comparators that is easy to interpret and helps to identify areas for improvement (Spiegelhalter 2005).

Benchmarking health system performance

Despite the methodological difficulties of comparative efforts, the diversity of benchmarking initiatives shows that national and regional health authorities are gaining increasingly from comparing their performance and learning policy lessons from better performers. The selection of benchmarks is becoming more pragmatic and increasingly is driven by the specific strategies of health systems and by their performance expectations. Performance measurement thus becomes the basis for policy discussions concerning how to improve health system performance and specifically about sharing how others have achieved higher performance in a particular context. For instance, a number of

Box 5.6.4 Benchmarking for better health system performance: example of the Commonwealth Fund in the United States

The Commonwealth Fund, a private organization in the United States, established the Commission on a High Performance Health System in 2005. This group of experts was assembled to analyse best practices from several health systems. Their benchmarking shows that Denmark performs better than any other country in Europe on measures of patient satisfaction and primary care; Germany is a leader in national hospital quality benchmarking; and the Netherlands and the United Kingdom lead on transparency in reporting quality data (Davis 2007).

Within the United States, the Commission also benchmarked states against each other across five key dimensions of health system performance – access, quality, avoidable hospital use and costs, equity, healthy lives (The Commonwealth Fund Commission on a High Performance Health System 2006). Cumulative and dimension-specific ranks were published along with an analysis of the policy implications. The results are publicly available and are intended to assist states to identify opportunities better to meet the population's health needs and learn from high-performing states (Cantor et al. 2007).

European countries have invested in efforts to benchmark their performance against countries such as Australia, Canada, New Zealand and the United States through the work of the Commonwealth Fund (Box 5.6.4).

In this perspective, a well-designed benchmarking system has the potential to guide policy development and can be used both prospectively and retrospectively (Nolte et al. 2006). It can support better understanding of past performance and the rationale behind certain performance patterns (retrospective use) and also help to revise strategies for improving future performance (prospective use).

Such strategy-based performance benchmarking systems have certain characteristics.

- Strategic focus: link between health system strategies and international benchmarking efforts ensures that policy lessons will be designed for those who can act upon the findings (the policy-makers).

- Adaptability and flexibility: benchmarking efforts can undertake both large (full health system comparisons) and narrower scope studies, using tools that can be administered in a time frame that matches the policy-makers' agendas (e.g. using patient survey comparisons such as that of the Commonwealth Fund).
- Data standardization: efforts are made to standardize data and facilitate credible comparisons.
- Policy focus rather than research focus: benchmarking systems are driven not by experts or researchers but by policy-makers supported by experts and researchers.
- Efforts to translate performance information and policy lessons for decision-makers: new tools (e.g. funnel plots) are used increasingly to represent performance information in rigorous yet explicit ways, conveying data in a meaningful manner while reducing the need to rank health systems in league tables.
- Sensitivity to political and contextual issues: interpretation of indicator data should not lose sight of the policy context within which they are measured; of the players involved in formulating and implementing policy; of the time lag needed to assess the impact of different policies; and of aspects of health care that remain unmeasured by available data.

Conclusions

This chapter reviews the reasons for increased governmental interest in international health system performance comparisons – they offer greater accountability and transparency and support strategy review and development. However, mutual learning is a third function that is becoming more important with the increasing scientific robustness of knowledge created through health systems research. Projects such as the OECD HCQI project or the Commonwealth Fund's cross-national benchmarking initiatives in the United States are two good examples of comparative efforts in this direction. The scope of experiences is growing and covers comparisons at different levels of the health system and from different perspectives. The methodological difficulties of such exercises can be classified and addressed over time but require investment from countries. Governments can achieve superior health system performance through the powerful policy instruments offered by linking performance measurement to performance management;

translating performance information in ways that are meaningful for policy-makers; and investing in benchmarking and mutual learning.

Finally, important requirements for fostering the value of international comparisons and their practical use for performance improvement are listed below.

- Recognize the value of information and make substantial investments in improving minimum data quality for developing and transition countries (e.g. through the Health Metrics Network) and data quality for developed countries (through projects such as the OECD HCQI).
- Build upon knowledge of how to resolve methodological issues in health system performance comparisons in order to strengthen such comparisons.
- Encourage international organizations to provide active support for data standardization efforts within their member states.
- Achieve a balance between process and outcome indicators in comparisons of health system performance in order to provide different but complementary insights into health-care processes.
- Avoid inconsistencies, strategic misalignment and (ultimately) health system sub-performance by selecting indicators that cascade across different (macro, meso, micro) levels of the health system through performance measurement and accountability mechanisms.
- Set up benchmark networks structured against common strategic objectives and performance patterns to build stronger analytical capacities within and between countries.
- Evaluate indicator data across countries with an adequate understanding of the regulatory and evaluative policies that underpin them.
- Develop and use graphic tools to convey performance information to policy-makers in a meaningful way.
- Undertake further research in health system performance management and share the results effectively among countries.

References

Agency for Healthcare Research and Quality (AHRQ) (2008a). *National healthcare disparities report 2007*. Rockville: US Department of Health and Human Services.

Agency for Healthcare Research and Quality (AHRQ) (2008b). *National healthcare quality report 2007*. Rockville: U.S. Department of Health and Human Services.

Arah, OA. Klazinga, NS. Delnoij, DM. ten Asbroek, AH. Custers, T (2003). 'Conceptual frameworks for health systems performance: a quest for effectiveness, quality, and improvement.' *International Journal for Quality in Health Care*, 15(5): 377–398.

Brand, H. et al. (2007). *Benchmarking Regional Health Management II (Ben RHM II) final report*. Bielefeld: Institute of Public Health of Nordrhein-Westfalen.

Cantor, JC. Schoen, C. Belloff, D. How, SKH. McCarthy, D. (2007). *Aiming higher: results from a state scorecard on health system performance*. New York: The Commonwealth Fund Commission on a High Performance Health System (http://www.commonwealthfund.org/publications/publications_show.htm?doc_id=494551).

Davis, K. (2007). *Learning from high performance health systems around the globe*. New York: The Commonwealth Fund. (http://www.commonwealthfund.org/publications/publications_show.htm?doc_id=441618).

Davis, K. Schoen, C. Schoenbaum, SC. Doty, MM. Holmgren, AL. Kriss, JL. Shea, KK (2007). *Mirror, mirror on the wall: an international update on the comparative performance of American health care*. New York: The Commonwealth Fund.

Drösler, S (2008). *Facilitating cross-national comparisons of indicators for patient safety at the national health-system level in the OECD countries*. Paris: Organisation for Economic Co-operation and Development (OECD Health Technical Papers no.19).

European Community Health Indicators Monitoring (ECHIM) (2008). *European Community Health Indicators Monitoring project*. Brussels: European Commission (http://www.echim.org/index.html).

European Health Interview Survey (EHIS) (2008). *European Health Interview Survey*. Brussels: European Commission (http://ec.europa.eu/health/ph_information/dissemination/reporting/ehss_01_en.htm).

Eurostat (2008). *Eurostat database*. Brussels: European Commission (http://epp.eurostat.ec.europa.eu/portal/page?_pageid=1090,30070682,1090_33076576&_dad=portal&_schema=PORTAL).

Feachem, R. Sekhri, NK. White, KI (2002). 'Getting more for their dollar: a comparison of the NHS with California's Kaiser Permanente.' *British Medical Journal*, 324(7330): 135–143.

FNORS (2007). *Project ISARE 3: Health indicators in the European regions*. Paris: Fédération Nationale des Observatoires Régionaux de la Santé.

Garcia-Armesto, S. Lapetra, MLG. Wei. L. Kelley, E. & Members of HCQI Expert Group (2007). *Health care quality indicators project 2006. Data collection update report*. Paris: Organisation for Economic Co-operation and Development (OECD Health Working Papers no. 29).

Goodspeed, SW (2006). 'Metrics help rural hospitals achieve world-class performance.' *Journal for Healthcare Quality*, 28(5): 28–32.

Health Results Team for Information Management (2006). *Strategy mapping*. Toronto: Ontario Ministry of Health and Long-Term Care (unpublished).

Himmelstein, DU. Woolhandler, S (2002). 'Getting more for their dollar: Kaiser v the NHS. Price adjustments falsify comparison.' *British Medical Journal*, 324(7349): 1332–1339.

Holland, WW (ed.) (1988). *The European community atlas of avoidable death*. Oxford: Oxford University Press.

Holland, WW (1990). 'Avoidable death as a measure of quality.' *Quality Assurance in Health Care*, 2(3-4): 227–233.

Hsiao, WC (1992). 'Comparing health care systems: what nations can learn from one another.' *Journal of Health Politics, Policy and Law*, 17(4): 613–636.

Hurtado, MP. Swift, EK. Corrigan, JM (eds.) (2001). *Envisioning the National Health Care Quality report*. Washington: National Academies Press.

Kaplan, R. Norton, D (1992). 'The balanced scorecard: measures that drive performance.' *Harvard Business Review*, 70(1): 71–79.

Kaplan, R. Norton, D (2000). 'Having trouble with your strategy? Then map it.' *Harvard Business Review*, 78(5): 167–176.

Kaplan, R. Norton, D (2005). 'Office of strategy management.' *Harvard Business Review*, 83(10): 72–80.

Kelley, E. Hurst, J (2006). *Health care quality indicators project conceptual framework paper*. Paris: Organisation for Economic Co-operation and Development (Health Working Papers no. 23).

Kerr, E. Fleming, B (2007). 'Making performance indicators work: experiences of US Veterans Health Administration.' *British Medical Journal*, 335(7627): 971–973.

Klein, R (1997). 'Learning from others: shall the last be the first?' *Journal of Health Politics, Policy and Law*, 22(5): 1267–1278.

Lavis, J (2006). 'Research, public policymaking, and knowledge-translation processes: Canadian efforts to build bridges.' *Journal of Continuing Education in the Health Professions*, 26(1): 37–45.

Legido-Quigley, H. McKee, M. Walshe, K. Suñol, R. Nolte, E. Klazinga, N (2008). 'How can quality of health care be safeguarded across the European Union?' *British Medical Journal*, 336(7650): 920–923.

Lousbergh, D. Buntinx, F. Geys, H. Du Bois, M. Dhollander, D. Molenberghs, G (2002). 'Prostate-specific antigen screening coverage and prostate cancer incidence rates in the Belgian province of Limburg in 1996–1998.' *European Journal of Cancer Prevention*, 11(6): 547–549.

Macinko, J. Starfield, B. Shi, L (2003). 'The contribution of primary care systems to health outcomes within Organisation for Economic Co-operation and Development (OECD) countries, 1970–1998.' *Health Services Research*, 38(3): 831–865.

Mattke, S. Epstein, AM. Leatherman, S (eds.) (2006). *International Journal for Quality in Health Care*, 18 (Suppl. 1): 1–56.

Mattke, S. Kelley, E. Scherer, P. Hurst, J. Lapetra, MLG. & HCQI Expert Group Members (2006). *Health care quality indicators project. Initial indicators report*. Paris: Organisation for Economic Co-operation and Development (OECD Health Working Papers No. 22).

McKee, M (2001). *The world health report 2000: advancing the debate*. Prepared for the European Regional Consultation on *The world health report 2000*, Copenhagen, Denmark, 3–4 September 2001. Copenhagen: WHO Regional Office for Europe (http://www.euro.who.int/document/obs/HSPAMcKeebackgrounddocument.pdf.).

Mountin, JW. Perrott, GS (1947). 'Health insurance programs and plans of western Europe: summary of observations.' *Public Health Report*, 62: 369–399.

Murry, CJL. Frenk, J (2000). *A framework for assessing the performance of health systems*. Geneva: World Health Organization.

Murray, CJL. Evans, D (eds.) (2003). *Health systems performance assessment. Debates, methods and empiricism*. Geneva: World Health Organization.

Nolte, E. Wait, S. McKee, M (2006). *Investing in health: benchmarking health systems*. London: The Nuffield Trust.

OECD (2001). *Health at a glance 2001*. Paris: Organisation for Economic Co-operation and Development.

OECD (2007). *Health at a glance 2007*. Paris: Organisation for Economic Co-operation and Development.

Rosenmöller, M. McKee, M. Baeten, R, Glinos, IA (2006). Patient mobility: the context and issues. In: Rosenmöller, M. McKee, M. Baeten, R (eds.) *Patient mobility in the European Union. Learning from experience*. Copenhagen: WHO Regional Office for Europe.

Rutstein, DD. Berenberg, W. Chalmers, TC. Child, CG. Fishman, AP. Perrin, EB (1976). 'Measuring the quality of medical care. A clinical method.' *New England Journal of Medicine*, 294(11): 582–588.

Schoen, C. Osborn, R. Doty, M. Bishop, M. Peugh, J. Murukutla, N (2007). 'Toward higher-performance health systems: adults' health care experiences in seven countries, 2007.' *Health Affairs*, 26(6): 717–734.

Schoen, C. Osborn, R. Huynh, PT. Doty, M. Zapert, K. Peugh, J. Davis, K (2005). 'Taking the pulse of health care systems: experiences of patients with health problems in six countries.' *Health Affairs (Millwood)*, 16(Suppl. web exclusives): 509–525.

Schröder-Bäck, P. (2007). *Results from the Benchmarking Regional Health Management II (BEN II) study* [PowerPoint presentation]. Bielefeld: Institute of Public Health of Nordrhein-Westfalen.

Spiegelhalter, DJ. (2005). 'Funnel plots for comparing institutional performance.' *Statistics in Medicine*, 24(8): 1185–1202.

Tawfik-Shukor, A. Klazinga, NS. Arah, OA (2007). 'Comparing health system performance assessment and management approaches in the Netherlands and Ontario, Canada.' *BMC Health Services Research*, 7: 25.

The Commonwealth Fund Commission on a High Performance Health System (2006). *Why not the best? Results from a national scorecard on U.S. health system performance* [online]. New York, The Commonwealth Fund (http://www.commonwealthfund.org/publications/publications_show.htm?doc_id=401577).

Veillard, J. Huymh, T. Kadandale, S. Ardal, S. Klazinga, N. Brown, A (2009). 'Making health system performance measurement useful to policy makers: aligning strategies, measurement and local health system accountability in Ontario.' *Healthcare Policy,* forthcoming.

Wait, S. Nolte, E (2005). 'Benchmarking health systems: trends, conceptual issues and future perspectives.' *Benchmarking: An International Journal*, 12(5): 436–448.

Westert, GP. Verkleij, H (eds.) (2006). *Dutch health care performance report 2006*. Bilthoven: National Institute for Public Health and the Environment.

WHO (2000). *The world health report 2000 – Health systems: improving performance*. Geneva: World Health Organization.

WHO (2007). *International classification of health interventions*. Geneva (http://www.who.int/classifications/ichi/en/).

WHO (2008). *European Health for All database* (HFA-DB) [online database]. Copenhagen: WHO Regional Office for Europe (http://www.euro.who.int/hfadb).

World Bank (1993). *World development report 1993: investing in health*. New York: Oxford University Press.

Zelman, WN. Pink, GH. Matthias, CB (2003). 'Use of the balanced scorecard in health care.' *Journal of Health Care Finance*, 29(4): 1–16.