Oxera

Agenda

Advancing economics in business

Who's in good health? Measuring performance in the NHS

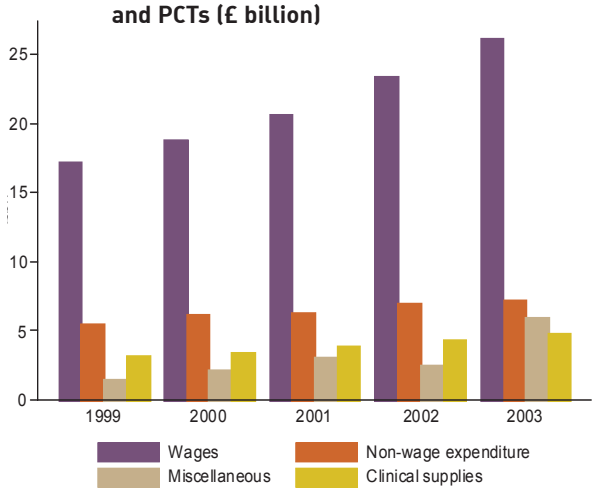
Public spending on healthcare has increased significantly over the past five years. One way to assess the benefits of such spending is to look at how well trusts can convert the increased resources into improved healthcare outcomes. This article reviews a number of suggested approaches to measuring performance in the UK's NHS and provides an illustrative analysis of the efficiency of the so-called acute trusts

When analysing the growth in NHS expenditure (see Figure 1), it is important to consider whether these increases in public spending have been accompanied by greater levels of output or higher-quality services. The Department of Health has shown that significant productivity improvements can be made to the delivery of public healthcare services:

- the NHS Purchasing and Supply Agency achieved savings of £228m on behalf of trusts in 2002–03, and a large portion of its bulk purchasing and logistics operations have recently been outsourced to DHL;[1]
- the NHS Institute for Innovation and Improvement showed that £78m could be saved by reducing spending on agency staff and improving labour productivity.[2]

The 2004 Gershon report on public sector efficiency also made a number of recommendations relating to how the NHS can improve its performance, such as jobs being moved out of London and the south-east where possible, back-office functions being shared between trusts, and procurement systems being combined to take advantage of the bargaining power of the NHS.[3]

Figure 1 Nominal expenditure for SHAs, trusts and PCTs (£ billion)



Source: Department of Health data.

However, despite increases in resource levels, some trusts have made staff redundant and are facing financial difficulties (see Table 1 below).

The correct response to this problem depends crucially on whether it is due to inefficiencies on the part of

NHS organisations explained

- ***Strategic health authorities*** **(SHAs) manage the local operations of the NHS on behalf of the Secretary of State for Health, and are the major link between individual trusts and the Department of Health. They are responsible for the development of plans for improving healthcare services in their local area and encouraging good performance of local NHS trusts. There were 28 SHAs covering England when they were created in 2002, but these have since been consolidated to ten.**
- ***Primary care trusts*** **(PCTs) provide 'first port of call' services such as GPs, dentists, opticians and pharmacists, along with NHS Direct. They control over 75% of the NHS budget and are responsible for the commissioning of mental health, ambulance, hospital and other services from related trusts in the local area.**
- ***Acute trusts*** **manage hospitals and have particular responsibilities for providing care and spending money efficiently. They also play a leading role in strategies for the development and improvement of hospital services.**
- ***Other trusts*****—remaining NHS services are provided by mental health trusts, ambulance trusts and care trusts. There is some overlap between the service types across the country—for example, some PCTs provide mental health services directly rather than commissioning a local mental health trust.**

individual trusts, over-provision of services, or inadequate funding. This article examines the ways in which performance in the health sector can be assessed and provides an example study, assessing the efficiency of acute trusts.

Table 1 Trusts with the largest deficits (2005/06)

	% of turnover	£m
Surrey and Sussex	25.5	40.8
Queen Mary's Sidcup	22.0	19.7
South Warwickshire General Hospitals	16.3	13.8
Queen Elizabeth Hospitals	14.4	19.2
Royal West Sussex	13.6	13.4

Source: Department of Health data.

What is productivity?

Productivity is traditionally considered to be the rate at which an organisation can transform its inputs into outputs: if more output can be obtained from a given level of existing inputs then productivity has improved (see Figure 2). This could either be because technology has improved, or because existing technology is employed at lower cost. Like firms, the NHS uses inputs (eg, nursing staff, drugs, beds) to create outputs (eg, heart transplants, GP consultations). Unlike some firms, however, the output of the NHS is multi-dimensional, sometimes unobservable, and difficult to measure accurately in its entirety.

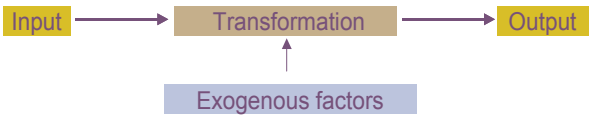
Performance assessments can aim either to minimise the amount of inputs used for a given level of output, or maximise the level of output achieved with a given level of inputs. Efficiency can be further broken down into technical efficiency (using the given inputs most effectively) and allocative efficiency (using the correct combination of inputs).

Another aspect of NHS performance is to choose the correct scale of operation. This is dependent on the costs of provision (is a hospital with twice as many beds twice as expensive to run?), and on how much society is willing to pay for different levels of healthcare. In the absence of a pure market mechanism, for social and distributional reasons, this is largely left to government and the political process to determine.

What are the appropriate inputs and outputs?

As with most industries, the health sector uses combinations of inputs to produce its outputs. These inputs can be grouped into labour (eg, medical and administrative staff), materials (eg, drugs) and capital (eg, hospitals and beds). Each trust will choose a combination of inputs which it believes will produce the best output for the communities served.

Figure 2 Assessing productivity



In assessing the performance of trusts it may be necessary to control for differences in the price of inputs if trusts have little control over them. For example, trusts in the south-east of England may have to pay higher wages to attract staff due to the higher cost of living.

Ideally, the output of a healthcare organisation would capture the value-added to individuals from a form of treatment, such as an increase in quality-adjusted life years. The main problem in observing this sort of measure is that there is little evidence on what would have been the health status of an individual without the treatment. In practice, it is often necessary to consider the activities undertaken by an organisation and adjust for the quality of the activity undertaken.

There is no one single output of the NHS in the way that a retail firm may consider its output the value of sales. Outputs do not consist simply of medical procedures—patients value the direct healthcare they receive, but also value attributes such as waiting times, cleanliness and friendliness.

Outputs can typically be divided into three groups:

- *measures of scale*—eg, number of operations, consultant episodes or patients;
- *measures of quality*—eg, waiting times, patient satisfaction;
- *measures of intensity and complexity*—eg, length of stay, type of treatment.

Ways of assessing performance

Given the complexity and diversity of activities undertaken, it is impossible to find one performance measurement system for the health service in its entirety. Instead, economists have tended to focus on either specific types of service (such as mental health trusts) or one aspect of the production function (such as labour productivity or consumer satisfaction). A number of the traditional methods of measuring performance can be applied to the NHS; however, there are problems inherent in each, and these require detailed consideration.

Key performance indicators

One of the simplest approaches is the use of key performance indicators (KPIs), which are *single* indicators of an organisation's performance (such as the number of finished consultant episodes per full-time member of medical staff). They have the advantage of having low costs of data collection and ease of

interpretation; however, they are only a partial measure of productivity since only a subset of activities can be included. Their major disadvantage if used for performance management is the distorted incentives they provide when considering only one input and one output of a more complicated process. As discussed above, the outputs of the NHS are numerous, and attempting to use only one piece of information is likely to cause problems when trusts divert resources from unmeasured to measured tasks, or focus only on certain activities to meet known indicators. Both types of behaviour may result in actual performance diverging markedly from that implied by KPIs. This problem can be alleviated somewhat by summarising multiple KPIs—for example, by employing data envelopment analysis (DEA).

The best-known measure of performance is the star rating provided by the Healthcare Commission. This indicator awards stars depending on how close each trust is to a specified target (such as the percentage of outpatients waiting longer than the standard 17-week target following a GP referral). While giving the public an easy-to-understand idea of how well their trust is performing, they do not explain why some trusts are particularly good at certain activities and poor at others. More importantly, star ratings do not consider the level of resource (input) being used to produce the output, or the level of demand for healthcare in each particular area. Since the key targets are set by the Healthcare Commission, and are designed to reflect the minimum standards that all organisations are expected to achieve, it is difficult to know if that is the best possible performance or whether a trust has done just enough to meet the target.

Labour productivity

As labour is the largest component of NHS costs, substantial effort has been made to understand how labour productivity is changing. Labour productivity indices give a good high-level view of this, but they focus on only one factor of production and give information relating to the average rate of productivity rather than what is possible.

In examining the results from labour productivity indices, it is important to understand what might drive differences in labour productivity, including local economic factors (such as labour shortages and cost of living) and demographic factors (such risk of disease). It is also worth considering whether labour productivity is limited by the capital available to the trust. Trusts with greater capital may achieve higher levels of labour productivity due to more modern facilities and technology.

Productivity indices

The current approach to assessing productivity adopted by the Office of National Statistics and the Department of Health is to calculate a weighted series of inputs and outputs and to compare growth in the two.[4] Output is measured by assigning value weights to various outputs and then adjusting for quality. This approach does consider the amount of resources used, but the resulting measures of productivity are crucially dependent on the weights employed, which are themselves of uncertain magnitude and may not be constant across all the trusts being compared. The same method is used to construct the input. Productivity will be said to have increased under this method if the rate of growth in inputs is less than the rate of growth in outputs. However, in a number of circumstances, this may be a misleading measure of NHS performance.

For example, if smaller hospitals are more expensive to run per patient than larger hospitals, the measure of output per unit input will increase as demand grows, even if there is no underlying change in efficiency. If the costs of delivering healthcare vary with the size of the delivery unit, this measure will have serious flaws in an environment where expenditure has been increasing rapidly, and where society's demand for healthcare is rising.

Frontier methods

Another branch of productivity measurement involves the estimation of a cost or production function and identifying the maximum possible performance for any particular business unit (eg, ward, trust or SHA). These methods produce a mathematical description of the technology and can be used to infer relative efficiency. Unlike KPIs or single factor measures of productivity, frontier methods can account for multiple factors of production (inputs) and the divergent benefits to society (outputs), allowing trusts to trade off the inputs and outputs available to them to find the optimum combination to meet the needs of the community.

Frontier methods can be used at the trust level and can take into account differences in characteristics between trusts, such as income or average age. The two main techniques in this area are stochastic frontier analysis (SFA)[5] and DEA.[6] Each has advantages and disadvantages. SFA is an econometric technique that requires assumptions about the relationship between inputs and outputs, and also requires significant amounts of data. SFA does, however, attempt to distinguish noise in the data from inefficiency such that all the unexplained differences in performance are not assumed to be inefficiency. DEA is a non-parametric approach that does not need a particular functional form. It has been used by regulators in other sectors, with adjustments to the inefficiency estimates to take account of unusual observations

An application of the DEA technique to acute trusts is presented below.

Data envelopment analysis

The standard DEA approach seeks to, first, compare one trust with trusts that are similar to it in terms of mix of outputs and scale size and, second, to attribute differences in outputs, controlling for inputs used, to inefficiency. DEA creates a space of feasible production units (ie, feasible trusts in this case) in terms of combinations of feasible output levels. The distance of a trust from the efficient boundary of this space, or frontier, is used as a measure of its efficiency. To compare units, DEA does not assume any particular functional relationship between inputs and outputs. Instead, it estimates the maximum possible output of a trust within the production space, when controlling for its mix of inputs to production. This provides the potential to measure the efficiency of a trust and to compare its efficiency with that of other trusts that are deemed to be efficient.

In creating the space of feasible production units, DEA needs to assume constant or variable returns to scale between costs (inputs) and the levels of the corresponding outputs. In general, an assumption of constant returns to scale technology can be used to estimate the efficiency savings only if either the relationship between inputs and outputs in the industry being modelled is indeed characterised by constant returns to scale, or if variable returns to scale hold, but operating units can be designed to exploit economies of scale and therefore reach the most productive scale size. If neither condition applies, variable returns to scale should be the basis for estimating the scope for efficiency savings. If changes in scale size cannot be effected, the efficiency savings estimated under constant returns to scale would not be realisable, as the operating units cannot reach the most productive scale size for reasons that are beyond managerial control.

DEA does not seek to split noise from inefficiency as SFA does. It is therefore appropriate to consider the application of further adjustments to take account of noise in the modelling. For example, some of the noise at the frontier can be accounted for by the removal of outliers at the efficiency frontier, with the effect of improving the overall efficiency estimate.

Case study: acute trusts

Using data on inputs and outputs for 159 acute trusts, it is possible to identify which trusts are operating efficiently. The input data is gathered from annual financial accounts,[7] and the outputs are from Hospital Episode Statistics.[8] The data is matched and cleaned to ensure consistency between inputs and outputs. The aim of the analysis is to identify efficient trusts and to generate targets for 'inefficient' trusts to catch up to best practice.

This example uses a multi-factor DEA model, as described in Table 2, and allows for variable returns to scale (ie, a 1% increase in output need not increase costs by 1%).

Inputs have been categorised into expenditures on wages, capital and materials (including medical supplies). No adjustments have been made to account for differential prices across trusts so, for example, it may be that trusts in the south-east are more efficient than these estimates would imply. The number of finished consultant episodes is used as a quantity measure of output, with quality and complexity being incorporated through data on waiting times and the proportions of old (60 years and over) and young (under 15 years) patients. There is scope to model both inputs and outputs more accurately by considering factors such as disease mix and income.

Table 2 DEA model for acute trusts

Inputs	Outputs
Labour	Finished consultant episodes
Capital stock	Mean waiting time
Capital flow	Mean length of stay
Materials	% of young or old patients

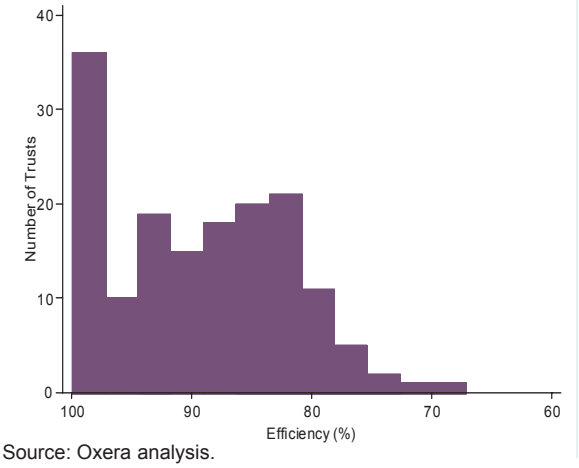
Source: Oxera analysis.

The results show that, of 159 trusts assessed, 29 were on the frontier and considered to be operating efficiently. Figure 3 below shows the distribution of efficiency across the acute trusts.

The average efficiency was 89.4% which implies that acute trusts could theoretically increase their outputs by an average of 10%. There may be reasons why the 10% increase in output is not possible for some trusts—for example, if wage costs are high in certain parts of the country, or if trusts specialise in more expensive types of treatment. Further work in this area could investigate whether there are other reasons for differences in performance.

The performance of the trusts highlighted in Table 1 as having the largest budget deficits is shown in Table 3. Two trusts have below-average efficiency, which may be contributing to their financial difficulty. However, two have above-average efficiency, and Royal West Sussex shows that financial problems can also be caused by other factors since it is rated as efficient relative to its peers.

Figure 4 Distribution of efficiency in acute trusts



Source: Oxera analysis.

Table 3 Trusts with the largest deficits and efficiency (2005/06)

	% of turnover	% efficiency (100 = benchmark)
Surrey and Sussex	25.5	89.09
Queen Mary's Sidcup	22.0	82.59
South Warwickshire General Hospitals	16.3	89.26
Queen Elizabeth Hospitals	14.4	81.07
Royal West Sussex	13.6	100

Source: Oxera calculations from Department of Health data.

How can performance assessment be used in healthcare services?

This example shows the sort of performance assessment that is possible using publicly available data. Such performance assessments have several advantages over traditional KPIs:

- the ability to handle multiple inputs and outputs;
- the ability to control for external factors that drive differences in performance;
- trusts with similar characteristics are compared;
- trust-specific targets based on the performance of peers are available;
- the presence of noise in the data is acknowledged;
- the impact of policy changes can be tested.

Performance assessments can also be linked to incentives for trusts to reduce costs and/or improve performance through performance-related pay and the setting of budgets. This allows for stronger incentives to be set based on achievable trust-specific targets.

Frontier-based methods of assessing performance also provide important information to policymakers, such as the extent of economies of scale and the optimal scale size for different types of trust. This sort of information is likely to be central to assessing the costs and benefits of possible mergers between trusts, and in understanding differences in productivity.

[1] Times Online (2006), 'American Firm is Hired to do all NHS Shopping', July 26th.
[2] NHS Institute for Innovation and Improvement (2006), 'Delivering Quality and value Focus on Productivity and Efficiency'.
[3] Gershon P. (2004), 'Releasing Resources for the Frontline: Independent Review of Public Sector Efficiency', HM Treasury.
[4] Office of National Statistics (2006), 'Public Service Productivity: Health, Economic Trends', **628**, March.
[5] Kumbhakar, S.C. and Knox Lovell, C.A. (2000), *Stochastic Frontier Analysis*, Cambridge University Press.
[6] Thannasoulis, E. (2001), *Introduction to the Theory and Application of Data Envelopment Analysis: A Foundation Text with Integrated Software*, Kluwer Academic Publishers.
[7] Department of Health, Financial Returns of Trusts, 2004/05.
[8] Hospital Episode Statistics, www.hesonline.nhs.uk.