Productivity

in the

National Health Service

by

P.E. Hart

July 2007

Centre for Institutional Performance
Department of Economics
University of Reading
Whiteknights
Reading
RG6 6AA

PRODUCTIVITY IN THE NATIONAL HEALTH SERVICE

by P.E. Hart

1. Introduction

To what extent are the many services provided by the NHS improving? Are the differences between the performances of different hospitals decreasing or increasing? Are the differences in NHS performance across regions in the UK being reduced? To answer such questions, and many others like them, it is necessary to measure the performance of the NHS. This is difficult and has led to an extensive and controversial literature on the subject. Recent examples are provided by Dawson *et al* (2005), Stevens *et al* (2006), and Castelli *et al* (2007), which also contain extensive references to earlier work. They focus on measures of productivity.

There is an even more extensive literature on the measurement of labour productivity and total factor productivity in the private sector. Section 2 of this note outlines an accounting approach to such measurement in the private sector and shows that the fundamental problems arising also hold for the NHS. Section 3 discusses the measurement of productivity in the NHS. Unlike firms in the private sector, the NHS does not charge for its services so that the output prices normally used to weight different services in an output index are zero. It is argued here that this fundamental problem is not overcome by the use of input prices in a cost weighted index of output. Hence indices of productivity based on cost-weighted indices of output should not be used to measure NHS total factor productivity.

2. The Private Sector

In the private sector there is a fundamental accounting identity which holds for each firm, for all firms together, in equilibrium and disequilibrium, in each accounting period:

$$(1) \qquad \sum_{i=1}^{h} q_i Y_i \equiv \sum_{j=1}^{k} p_j X_j , \qquad \begin{array}{l} i = 1,2,...,h \\ j = 1,2,...,k \end{array}$$

where q_i is the price of the ith output Y_i of the typical multi-product firm and p_j is the price of the jth input X_j. That is, the trading account of a firm is summarised by (1) and the identity holds because profit, denoted by p_k X_k, is the positive or negative residual which brings the account into balance. X_k may be regarded as capital input and p_k is its rate of return.

The profit, p_k X_k, is gross and includes stock appreciation, depreciation and taxation. Each item could be specified separately in a disaggregation of p_k X_k, but the identity in (1) would still hold for each firm. It would also hold for the aggregate of all firms, though if an estimate of aggregate output were required, the extensive double counting would have to be removed since the output of one firm may be the input of another.

Adjustments would also have to be made to allow for the different accounting years of different firms.

The identity in (1) holds in equilibrium and disequilibrium, in times of inflation and deflation, in conditions of high employment and unemployment, in perfect and in imperfect competition. It is certainly not assumed that q_i reflects the marginal social value of Y_i. Moreover, the quality of each Y_i may vary over time and such changes are not usually captured in the statistics. In addition, in our dynamic world, some new outputs are always being introduced, while others are phased out.

Obviously, no index of total productivity could be compiled using the ratio of price weighted outputs on the LHS of (1) to the cost-weighted inputs on the RHS of (1) because the index would always be unity, or 100 in index terminology. We could omit $p_k X_k$ from the RHS of (1) to remove the identity and compare the profit or rate of return in different firms, which would be a useful guide to different performance.

We could use total factor productivity which increases when the increases in the outputs, Y_i, exceed the increases in the inputs, X_j. Since the identity in (1) must be preserved, the increase in total factor productivity must be offset by the increase in input prices, p_j, exceeding the increase in output prices q_i. There is a dual relationship between the growth of total factor productivity and the growth of the price differential between inputs and outputs, as emphasised by Jorgenson and Griliches (1967). For example, the increase in total factor productivity might be absorbed by a large increase in profitability, p_k, or by increases in wages as the result of trade union pressure. Both results would yield increases in p_j exceeding increases in q_i. If total factor productivity decreases, it is likely that the fall in profitability, p_k reflects the excess of the q_i over the p_j.

The RHS of (1) may be simplified and decomposed into labour, capital and other costs (eg. raw materials) to give:

(2) $qY \equiv wL + rK + pM$

where $qY = \Sigma q_i Y_i$ is total sales, wL is total labour cost , rK is profit, and pM is total raw material and other costs. Real output may be written:

(3) $Y \equiv (w/q)L + (r/q)K + (p/q)M$

and labour productivity is given by:

(4) $Y/L \equiv w/q + (r/q)K/L + (p/q)M/L$

In (3) and (4) the standard official practice of using single deflation is followed with the output price index q acting as the deflator. Clearly, increases in labour productivity are related to many variables in addition to labour input, namely w/q , r/q , K/L, (p/q) and M/L .

Suppose firms have data on p_j, X_j and Y_i but all data on q_i are lost in their computers. Their cost accountants then estimate the full cost price of each Y_i, including a profit mark-up, by allocating the known total costs to each output. Denote this estimate by c_i. If the cost accounting is 100% accurate $c_i = q_i$. Once again the ratio of cost-weighted outputs to cost-weighted inputs would be unity and cannot be used to measure total productivity. In practice, cost accounting is not perfect and differences in this ratio would be observed. Such differences would reflect errors and omission in cost accounting rather than differences in the performance of different firms

The cost weighted index of output, $\Sigma\, c_i\, Y_i$, could be used to measure labour productivity by dividing it by some measure of L. However, we would not know whether any observed differences result from differences in cost accounting methods or from real differences in labour productivity. The allocation of total costs to each output is very difficult, sometimes even arbitrary, and errors are bound to arise. Moreover, as indicated by (3) and (4), labour productivity is a partial measure and, as a result of the other variables on the RHS, may be a misleading guide to the performance of firms.

3. The NHS

In the NHS the q_i are zero, apart from the prices of minor services such as car parking and refreshments for visitors. To overcome this problem, cost weighted indices of output are calculated from data on p_j X_j which may be written $\Sigma_i\, c_i\, Y_i$. If the c_i are calculated correctly the identity in (1) becomes:

$$(5) \qquad O_c \equiv \sum_i c_i Y_i \equiv \sum_j p_j X_j$$

and the ratio of the LHS to the RHS is still unity and cannot be used to measure total productivity. It is very difficult to measure total factor productivity. The q_i are zero so the share of each Y_i in total output cannot be measured. The positive c_i substitutes for q_i are based on p_j so the dual price differentials will be small, reflecting difference in input and output weights and the inevitable differences in cost accounting methods. We must not be surprised when Dawson *et al* (2005) estimate that the growth of total factor productivity in the NHS and in hospitals was slightly negative between 1998 and 2004.

The measurement of $\Sigma\, Y_i$ in the NHS is also extremely difficult. The numbers of categories of NHS activities (eg operations, diagnostic tests, consultations etc.) are available and the cost shares of each sector are estimated. Such activities contribute to the ultimate output, Y_i, which is the reduction in the disutility of a patient. Standard measures of productivity in manufacturing industry are cardinal and provide information on the distance between the performance of firms. For example, the labour productivity of firm A is 10 per cent higher than in firm B. Cardinal measures of utility, and hence of disutility, are ruled out by economists, but in principle ordinal measures of disutility may be made. That is, in principle, we may be able to rank disutilities even if we cannot measure them.

In practice it is difficult to obtain agreement on ordinal measures of health care outcomes. A common method is to use quality adjusted life years or QALYs. The extra years of life resulting from an activity is a cardinal measure, though it is subject to considerable errors of estimation. The quality of life in these extra years is ranked by a subjective scale derived from people's opinions. Health professionals, patients and the general public may rank health problems differently. Even setting this scale to zero at death will not command universal agreement. Some patients with dreadful health problems may regard death as preferable so their own measure of QALY is negative and outside the scale. Other patients with serious disabilities may learn to adapt to them, to the surprise of health professionals who have assigned them a low QALY.

Even if a consensus on the ranking of qualities of life can be reached, the resulting QALY is still ordinal not cardinal. Hence calculating £ cost/QALY, weighting by QALYs, discounting QALYs etc. amounts to pretending that ordinal measures are cardinal, which may not be appropriate. Of course, the use of ranks as instrumental variables has a long history in applied economics and we may well have more confidence in the rank of a health benefit than in any estimate of its amount.

Another problem is that inappropriate measures of Y_i do more harm than good because they stimulate sub-optimal practices of health professionals known as gaming. For example, if an official health benefit target is to reduce the numbers of patients on waiting lists for elective surgery, it can be achieved by concentrating on the less difficult and shorter operations (eg hernias rather than heart transplants) even though the patients are less seriously ill. Again, if the target is to reduce the average waiting time for patients requiring non-elective surgery in accident and emergency units (A&E), the less serious cases can be left in the care of paramedics in ambulances parked outside the hospital so that the resulting delay in admissions reduces the average waiting time in A&E. Waiting times are regarded as important components of performance measures. Chief executives of hospitals, who are responsible for allocating resources, lose their jobs if their hospitals have low scores (ie no "stars") in the official league tables of hospital performance. In such circumstances, it is not surprising that they chase spurious output targets. Perhaps the measures of input costs, to which we now turn, are more reliable.

Major difficulties arise in estimating and allocating capital costs in the NHS. Data on the depreciation of past capital expenditure are available in the Trust Financial Returns. Capital services provided by capital purchases in the current year (eg computers, software, medical equipment) are assumed to be one third of such capital expenditure. The share of capital services in total NHS inputs is about 8% compared with labour's share of 72% and 20% for intermediate inputs. Nevertheless differences in the accuracy of estimating capital inputs into each NHS unit would affect comparisons of productivity. Estimates of intermediate inputs (drugs etc.) are also taken from the Trust Financial Returns and are equivalent to p M in (2).

Labour is the most important input and its skill content has increased over the years .It may be measured indirectly by deflating the total wage bill by a suitable index of wages or directly by counting numbers of full-time equivalent employees or their total hours worked. Dividing a cost weighted output index by an index of labour input yields an index of labour productivity. As noted above, any differences observed could result from other variables in (4) and also from differences in cost accounting.

The paper by Stevens *et al* (2006) estimates labour productivity in the NHS by dividing a cost weighted output index by a weighted average of staff types given by:

$$(6) \qquad L = \sum_m n_m w_m = \sum_m n_m \left(n_m \overline{w}_m \right) / \sum_m n_m \overline{w}_m$$

where n_m is the number of staff of type m, w_m is their share of the total NHS wage bill and $\overline{w}_m$ is the average wage of type m staff.

The total number of staff, $N = \sum n_m$ is not used to measure L. The different hours, skills, and qualifications produce different amounts and qualities of labour input which affect L but do not affect N. Note that the large weight given to relatively small numbers of people at the top of the NHS hierarchy, and the small weights attached to the large numbers of people at the bottom, tends to reduce L below N. This has the effect of increasing labour productivity.

This measure of labour input differs from the standard direct and indirect methods described above. If N is constant, but there is a transfer of staff from lower paid to higher paid grades, as a result of general increases in skills, then L increases. This might reduce labour productivity if the resulting percentage increase in O_c is smaller. If the increase in O_c accurately reflects the increased labour costs, then labour productivity does not change, in spite of the increased skill content, because both O_c and L increase in the same proportion. In practice, different hospitals would be likely to have different changes in labour productivity as the result of differences in cost accounting. Such estimates may well be a misleading guide to policy makers.

4. Conclusions and Policy

Technological progress in medicine is rapid. The new drugs and procedures tend to be expensive, so medical costs increase more quickly than the general price level. Longevity is also increasing, partly as the result of the costly improvements in medicine. Faced with spending even more billions of pounds on the NHS, the Department of Health tries to ensure that the taxpayer obtains value for money. To do this, it needs reliable measures of the performance of the NHS and commissions large teams of high-class researchers to develop them. The costs of inputs can be measured, but the required cardinal measures of the benefits of the outputs cannot be made. At best, ordinal measures of health benefits, such as QALYs, can be developed, but these are subjective and hence controversial.

Above all, there are no prices for NHS outputs because they are free to NHS patients at the point of delivery. Thus NHS outputs cannot be weighted by their prices. Instead, they are weighted by their costs, after some transformation. Clearly, measuring changes in total factor productivity by comparing changes in costs with changes in transformed costs (shadow prices) is not very helpful. In fact it may be harmful: the morale of hard-working health professionals may be damaged when they are told that the growth of total factor productivity in the NHS is negative.

Introducing positive q_i in (1) would enable measures of performance to be made, but would require a fundamental change in policy, which might not be feasible politically. However, there are signs of change. In many parts of the UK, dentists have left the NHS and so their q_i are determined in the market. In optometry, most adult patients have to pay for their eye examinations though free NHS examinations are provided for children, the elderly and those poor enough to be on passport benefits. Furthermore, in an interesting development, patients on passport benefits are given vouchers, which cover the cost of budget spectacles (equivalent to the old free NHS spectacles) and which they are allowed to supplement, if they wish to buy more fashionable (and hence more expensive) frames, ie they make co-payments.

If positive q_i were ever introduced into the mainstream NHS, the above exemptions for children etc. could be maintained and monitored by hospital almoners, as in the days before the NHS. Even if other patients were not prepared to pay the full q_i they might accept the principle of co-payments. That is, they would make a direct personal contribution towards the market determined q_i, in addition to their indirect contribution through taxation. Private health insurance would be available to cover the risk of having to make a co-payment.

Would this hybrid NHS financed by co-payments and taxation be politically feasible? It is worth seeking the views of the electorate. The alternative is to continue dealing with the NHS in its present form and accept the inappropriate performance measures, spurious targets, gaming and arbitrary allocations of resources of an Alice in Wonderland world. Even the rich, with their private health insurance, have to use the Accident and Emergency units of NHS hospitals.

Department of Economics
University of Reading Business School
(July 2007)

References

Castelli, A., Dawson, D., Gravelle, H., Jacobs, R., Kind, P., Loveridge, P., Martin, S., O'Mahony, M., Stevens, P., Stokes, L., Street, A. and Weale, M. (2007) 'A new approach to measuring health system output and productivity'. *National Institute Economic Review*. April.

Dawson, D., Gravelle, H., O'Mahony, M., Street, A., Weale, M., Castelli, A., Jacobs, R., Kind, P., Loveridge, P., Martin, S., Stevens, P. and Stokes, L. (2005) 'Developing new approaches to measuring NHS outputs and productivity: final report'. *York Centre for Health Economics*. Research Paper 6.

Jorgenson, D.W. and Griliches, Z. (1967) 'The explanation of productivity change'. *Review of Economic Studies*, Vol. 34, 249-281.

Stevens, P., Stokes, L. and O'Mahony, M. (2006) 'Metrics, targets and performance'. *National Institute Economic Review*, July.