OPI-ESRC Seminar Series on Health Services Productivity

An Introduction to Measuring Efficiency and Productivity in Health and Health Care: Andrew Street, University of York, UK & Bruce Hollingsworth, University of Monash, Australia;
with the Centre for Health Management, Imperial College, London
12 and 13 January 2004

Summary

The measurement of efficiency and productivity in the delivery of health services is attracting increasing attention as public sector policy makers attempt to gain the most from limited health budgets by introducing incentives to encourage managers to innovate, and real or quasi-market forces to promote competition between production units. This has stimulated advances in techniques to analyse productivity, most recently, Stochastic Frontier Analysis and Data Envelopment Analysis. These advances have been accompanied by an increased awareness of the costs of making policy errors due to an inadequate understanding of the most important drivers of healthcare productivity improvement. Dangers arise from the blind conversion of productivity results into rewards and sanctions and from the misspecification of the relationships between health services and outcomes.

Background

Increasing pressures on health sector resources have stimulated interest over the last decade in health services productivity and in ways to improve it. In OECD countries, with large publicly funded healthcare sectors, the main interest in productivity has focussed on the use of productivity measures to manage rewards and penalties that are intended to encourage efficiency-seeking managerial behaviour. Elsewhere, the shift in focus towards the public purchase of healthcare from private providers is increasing interest in the relative efficiencies of public and private service provision. On a global level, international health agencies are interested both in productivity comparisons between countries and also in developments that are geared towards the more effective use of limited resources. In summary, analysts have two kinds of interest in measuring health service efficiency: in cross-sectional comparisons between healthcare production units within areas or countries, or between countries; and in variations in productivity over time. Based on these interests, the aim of this seminar was to review recent advances in productivity measurement, with a particular focus on their applications in the management of performance-based purchasing.

The measurement of health service efficiency and productivity

In principle, productivity measures may be either output-oriented, where the independent variable of the production function is a measure of health outcome or health service activity; or cost-oriented where the independent variable is a measure of health service cost. These are `duals' but only under the restricted condition of constant scale economies. The choice of explanatory variables is also dependent on the analyst's perspective. For example, a technical approach, which is based of the theory of the firm, focuses on the impact on output of the level and mix of inputs and is concerned with the maximum output that can be obtained from different combinations of production factors. The Cobb-Douglas function is a typical formulation of a production function based on this perspective. Alternatively, the analyst may wish to employ a behavioural approach, which focuses on organisational effort and managerial efficiencies. Agency theory has been useful in helping to clarify the incentives that managers have to improve efficiencies, given the existence of asymmetric information between them (the agents) and their principals (the regulator, commissioner or consumer). Production functions based on this perspective typically relate output maximisation to the cost of effort and transactions costs, as well as to the costs of factors of production. The UK NHS has used a variety of measures to estimate efficiency. These include the NHS Efficiency Index, the NHS Performance Indicators and, production function estimates.[1]

Analytical tools

Until recently, Ordinary Least Squares (OLS), or some derivative of it, has been the standard econometric method used to estimate production (or cost) functions. OLS generates a line of best fit through a set of data points which, according to standard econometric analysis, regards the residual to be the result of random influences and of measurement error. Applied to output oriented productivity data, points lying above the line of best fit represent units performing above average and points lying below the line of best fit are units performing below average. The opposite interpretation applies in the case of cost functions.

Fig 1: OLS & Corrected OLS

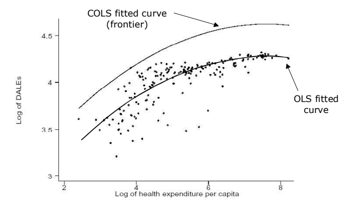


Figure 4: Frontier Production Function for 191 countries 1997: Corrected Ordinary Least Squares

CENTRE FOR HEALTH ECONOMICS THE UNIVERSITY of York

A development of OLS, Corrected Ordinary Least Squares (COLS), estimates a shift variable in order to place the line of best fit through the best performing units (Fig 1). In some sense, this line represents the production frontier, although it is only defined by the performance of the production units under analysis. The distance between any individual data point and the COLS line represents the extent to which performance falls below those of the best observations in the sample. Relative efficiency is measured by the ratio of this distance to the distance between the data point and the axis.

1. Street, Andrew (2000) Confident about Efficiency Measurement in the NHS? Healthcare UK reviews the cost indices used by the UK NHS.

More recently, two new analytical tools have been developed that improve upon both OLS and COLS: *Stochastic Frontier Analysis* (SFA) (a derivative of OLS) and *Data Envelopment Analysis* (DEA). These are now regarded as the most advanced techniques for measuring relative efficiency. SFA (Fig 2) improves upon COLS by partitioning the residual between a true error component and an inefficiency component. The distance between any individual data point and the fitted line represents the extent to which performance falls below the optimal observations. Relative efficiency is measured by the ratio of this distance to the distance between the data point and the axis.

Fig 2: Stochastic Frontier Analysis

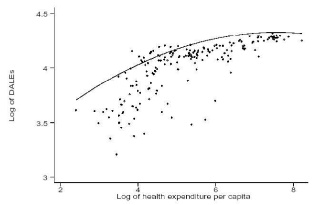


Figure 5: Frontier Production Function for 191 countries 1997: Stochastic Frontier, Truncated Normal

CENTRE FOR HEALTH ECONOMICS THE UNIVERSITY of York

DEA (Fig 3) plots the extreme data points in the sample, and hence `envelops' the available data. It uses an optimisation programme to calculate relative efficiency based on the allocation of a weighting to each input and output in the analysis. Unlike SFA, DEA provides information on the changes that can be made to inputs and outputs in order to maximise efficiency, that is, to move onto the frontier. Its great advantage is that it can handle multiple outputs as well as multiple inputs.

Fig 3: Data Envelopment Analysis

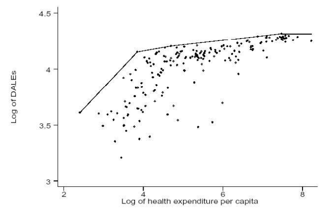


Figure 3: Frontier Production Function for 191 countries 1997: Data Envelopment Analysis

CENTRE FOR HEALTH ECONOMICS THE UNIVERSITY of York

These techniques differ in a number of ways. The first relates to their theoretical basis. SFA is *parametric* and appeals to economic theory. It estimates the combinations of inputs that generate the greatest volume of output, in the case of an output-oriented production function, or, the case of a cost-orientated function, the least-cost combination of inputs required to generate a given level of output. DEA is *non-parametric* since it is exclusively moulded by the data. The production frontier is defined in terms of the performance of a peer group of best performing units.

Second, the two techniques differ in their interpretation of the distance from the frontier. In the case of SFA, an allowance is made for 'error', while in the non-parametric frontier all shortfalls in performance are attributed to inefficiency. This difference arises from the way they use information: full sample information in the case of SFA and closest peers in the case of DEA.

SFA and DEA: how do they compare?

SFA	DEA
Parametric	Non-parametric
PF defined by theory	PF defined by the data
Based on modelled relationship between inputs and outputs	Based on selected outputs
Output weights assigned by the analysis	Output weights vary freely
Can only handle one output	Can handle multiple inputs and outputs
Relative efficiency defined by distance from the frontier	Relative efficiency defined by distance from the nearest peer
Results may not be easy to interpret intuitively	More flexible and user-friendly

Third, the two techniques differ in the source of weights assigned to each output.

Ideally, outputs should be weighted to reflect their social value. In the case of SFA, weights are generated by the estimation and are equal to the mean marginal cost of output in the sample. This implies that expenditure choices reflect social values. In contrast, DEA allows weights to vary freely. This means that although each organisational unit is evaluated in the best possible light, it is inappropriate to rank either DEA efficiency scores or the efficiency of public organisations based on these scores.

Methodological difficulties and their consequences in practice

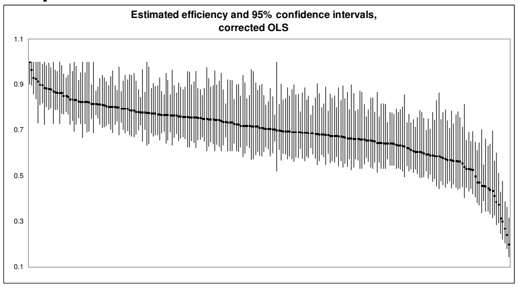
The assessment of health services productivity is constrained by two main factors. First, health care represents a diverse bundle of services and second, it is methodologically complex to measure health service inputs and outputs accurately. These difficulties influence methodologies at different levels. First, there is a general difficulty arising from the fact that efficiency is unobservable. Measurable proxies are therefore required, although these may only partially reflect the true relationships between outputs and inputs. In terms of output-oriented measures, the analyst has a broad choice between measuring some proxy of health outcomes or the volume and quality of health services being delivered. For example, on a global level, WHO's analysis of health system performance used Disability-Adjusted Life Expectancy (DALE) as a proxy for health outcomes.[2] This is immediately problematic because it is difficult to measure morbidity and mortality accurately in many countries and also because it is difficult to estimate population size. In contrast, most output measures in the UK are related to service volume and quality. The complexity of the output data is generally reduced by classifying services into Health Related Groups (HRGs). In practice, the results have been problematic because of the general focus on acute care cases, at the exclusion of both long stay patients in hospitals and primary care consultations. In addition, NHS measures have also typically failed to consider the individual characteristics of different hospitals. For example, although some hospitals may appear to be relatively inefficient, this may be because resources are held as spare capacity in an attempt to increase access and the timeliness of treatment. As a result inefficiency can be falsely attributed. This may be problematic at a policy level, since it may be possible for productivity to appear to be constant despite increases in funding. In this regard, it is necessary to measure all factors that are included in the national policy framework. Only this will ensure that increased policy demands, such as decreased waiting times and increased primary care consultations, are included in productivity measures.

With regard to input-oriented measures, there exists an even greater choice of variables that might be associated with higher levels of output. For example, the explanatory variables chosen by WHO's global study were education and expenditure on health services, although a host of other factors may also contribute to DALEs, not least, household income. The health production function of individual countries based upon the WHO model can therefore be regarded as fairly arbitrary. It is noteworthy that production functions, that are designed to explain levels of service activity, generally use service costs as the main explanatory variables. These are usually defined by the costs of labour and capital. However, this has been challenged on the grounds that, while they may be related to service volume and quality, environmental factors such as prices in local factor markets may be neglected.

---

2. WHO (2001) The World Health Report 2000 Health Systems: Improving Performance

Fig 4: Errors associated with relative efficiency estimates for a group of NHS hospitals



Even if some degree of consensus can be achieved on the structure of the production function, problems may arise both in validating the accuracy of the data and in actually extracting it. In addition, the accurate apportionment of inputs to outputs must be regarded as a crucial element in determining relative efficiency, since a failure to take this into account can provide a major source of estimation errors. Measurement and estimation errors may have particularly serious consequences if relative productivity measures are used by regulators to allocate rewards and penalties between facilities that are performing within a relatively narrow range. This is illustrated by the confidence limits associated with productivity estimates from a group of UK NHS hospitals (Figure 4), which reveals that significant differences only exist at the extreme ends of the performance range.

Conclusions

Although substantial advances have been made in productivity analysis in recent years, the effective use of productivity measures is dependent on the consideration of a host of factors that may influence organisational performance. These include differences in the target population, in the external environment, in the quality and cost of available resources, in accounting practices and on organisational priorities, as well as differences in efficiency. In this regard, it is noteworthy that DEA and SFA can both account for environmental factors. DEA can include environmental variables either as inputs, albeit at the risk of labelling any given organisation as fully efficient, or in the second stage of a two-stage OLS analysis, although the inclusion of environmental variables in time series data analysis may lead to serially correlated efficiency scores. In the case of SFA, all regressors should be environmental variables, with those that test statistically significant explaining some random error if the model was previously misspecified and/or some inefficiency if the error terms are correlated.

Apart from these factors, the interpretation of productivity analysis also needs to take the trade-offs managers consider when allocating resources between competing objectives and priorities. In addition, organisations operate in an historical context. Thus, while endowments of investments and past efforts may affect current performance, current investments are intended to influence future attainment. The analysis of panel data may be more appropriate than cross-sectional performance analysis since it provides a better measure of the effects of investments on performance.

For all of these reasons, changes in incentives or recommendations for either input minimisation or output maximisation should not be made blindly. Rather, they should be advanced with caution and reconciled with managers' priorities. Indeed, analysts should apply techniques for measuring productivity with an understanding of the full range of factors influencing performance. At the same time, every possible attempt should be made to develop a coherent model of production, in which the results are interpreted as part of a broader portfolio of performance indicators. This should consider not only the input and output variables that are included in the analysis, but also those that are excluded from the study.

It is noteworthy that there are no *a priori* reasons for selecting one analytical technique over another, since the main determinant of the technique to be used should be the purpose of the analysis and the nature of the data. It is also important to define the theoretical framework underpinning the model and the reasons chosen for model specification. Traditional modelling criteria, for example, may be inappropriate if the aim is to infer individual efficiencies, since statistically insignificant variables may be significant determinants of efficiency.

Finally, it is imperative that the costs of incorrect inferences be made clear. In this regard, an estimate of confidence limits may be useful since it allows productivity measures to be interpreted more cautiously and hence reduces the damage costs of naïve interpretations. This points to the crucial importance of improved data collection since, although additional and more accurate data comes at a cost, it can help to improve the precision of estimations by reducing the scope for measurement error.

Rianna Lisa Mohammed
University of Cambridge
February 2004

Citation: Mohammed, Rianna (ed) 2004 An Introduction to Measuring Efficiency and Productivity in Health and Health Care: OPI-ESRC Seminar Series on Health Sector Productivity given by Andrew Street, University of York, UK & Bruce Hollingsworth, University of Monash, Australia at the Centre for Health Management, Imperial College, London: Oxford Policy Institute.

Acknowledgments: The figures are taken from Andrew Street's presentation at the workshop. Further details about the seminar contents can be obtained by contacting
Andrew Street
ads6@york.ac.uk or
Bruce Hollingsworth
bruce.hollingsworth@buseco.monash.edu.au.

Further reading:

1. Cooper, W., Seiford, L.M., Tone, K. *Data Envelopment Analysis*. Kluwer: Dordrecht, 2000.
2. Coelli, Tim (1996) A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimatio. *CEPA Working Paper 96/07* Centre for Efficiency and Productivity Analysis, University of New England http://www.uq.edu.au/economics/cepa/software.htm
3. Coelli, T., Rao, D.S.P., Battese, G.E. *An introduction to efficiency and productivity analysis.* Kluwer: London, 1998.
4. Hollingsworth, B. and Parkin, D. Efficiency and Productivity Change in the English National Health Service: Can Data Envelopment Analysis provide a robust and useful measure? *Journal of Health Services Research and Policy;* 2003: 8 230-236.
5. Hollingsworth, B. Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science*; 2003: 6(4) 203-218.
6. Hollingsworth, B. and Wildman, J. The Efficiency of Health Production: Re-estimating the WHO Panel Data using Parametric and Nonparametric Approaches to Provide Additional Information. *Health Economics;* 2003: 12(6) 493-504.

7. Street A. How much confidence should we place in efficiency estimates? *Health Economics*, 2003; 12: 895-907.
8. Hollingsworth, B. Data Envelopment Analysis and Productivity Analysis: a review of the options. *Economic Journal;* 1999: 109(456) 458-462.
9. Gravelle H, Jacobs R, Jones AM, Street A. Comparing the efficiency of national health systems: a sensitivity analysis of the WHO approach. *Applied Health Economics and Health Policy*, 2003: 2(3): 141-147.
10. Software can be downloaded from http://www.uq.edu.au/economics/cepa/software.htm