Measuring efficiency to inform health policy

Peter C Smith, Andrew Street
Centre for Health Economics
University of York

Paper presented at the Health Economists' Study Group, 8-10 January 2002
Nuffield Institute for Health, University of Leeds
Department of Health

Work in progress
Please do not quote without permission of the authors

Abstract

Policy makers are increasingly seeking to use 'off-the-shelf' analytical tools such as data envelopment analysis (DEA) and econometric models to draw conclusions about the performance of individual health systems and organisations. The most prominent example is the rating of health system performance by the World Health Organisation, but numerous other applications have been reported. The analytical sophistication of such methods has reached an advanced stage of development. In applying DEA, analysts can explore different scaling assumptions, partition estimates to measure the form that inefficiency takes, and bootstrap estimates to assess statistical significance. Econometric techniques allow different functional forms, different distributions of inefficiency, and the calculation of confidence intervals around inefficiency estimates. However, such approaches rarely yield definitive or consistent conclusions. Quite modest changes in the choice of analytic technique and model specification can lead to major changes in inference about efficiency. In this paper we provide an assessment of the assumptions underlying the various methods, highlighting their strengths and limitations and their consistency with policy requirements. For the health sector, we make recommendations as to the pre-requisites for undertaking efficiency analysis, the analytical process that should be followed, and the interpretation that should be placed on the results.

Measuring efficiency to inform health policy

Peter C Smith, Andrew Street

1 Introduction

Extraordinary progress has been made in the theory and empirical estimation of models of productivity and efficiency over the last 25 years. Advances in econometric methodology, an explosion in data provision, and the ready availability of high quality software have all contributed to the prodigious growth of interest in the topic. Moreover – and unusually for such complex analytic technology – the methods are increasingly being considered by policy makers as a tool for influencing real economic behaviour.

The purpose of this paper is to examine whether or not the methodologies are ready for such policy use. The conclusion will be that – for relatively simple, well understood production processes, supported with good data, such as the water industry – the techniques can be valuable. However, we conclude that for complex, poorly understood processes, such as those undertaken by police forces and health systems, the methods must be used with great caution.

In the next section we outline reasons for the policy interest in efficiency measurement. this is followed by a brief overview of the main empirical techniques that are used to generate overall measures of organisational efficiency. Section fours draws attention to the following areas of controversy:

- output weights: how should trade-offs between competing objectives be considered?
- model construction: how accurately is it possible to model the underlying production process?
- exogenous factors: what environmental factors should be taken into account when comparing performance?
- dynamic effects: to what extent are past inheritances and investments in the future relevant to the assessment of current performance?

We conclude with general recommendations as to how to proceed with this line of investigation.

2 What are the objectives of policy makers?

Broadly speaking, indications of industry-wide and organisational efficiency can be inferred from two types of data:

- Performance indicators, which measure specific factors thought to provide a partial reflection of underlying efficiency.
- Comprehensive measures, designed to provide an indication of overall organisational efficiency.

It has been traditional in the UK public sector to take a piecemeal approach to measuring and monitoring efficiency, relying on selected performance indicators that capture aspects of performance related to efficiency. Performance indicators have much to recommend them. They are focused on specific aspects of performance, are readily easily measured and validated, and are easy to interpret (in isolation, at least). However, there are two major drawbacks in using individual performance indicators. First, they provide only a partial indication of performance or efficiency.

Second, they may provide conflicting messages: an organisation that appears to do well on one indicator may perform less successfully when considering another. Given that they are partial and may produce conflicting messages, it is not straightforward to draw conclusions about overall organisational performance from a narrow range of indicators.

In order to rectify this situation, policy-makers have been considering the possibility of devising comprehensive measures of performance, which seek to assess the overall success with which individual organisations are meeting objectives. This approach has appeal for various reasons:

- In contrast to piecemeal examination of single performance indicators, global indices of efficiency can offer a rounded assessment of system performance. This is particularly important when inputs (in the form of expenditure) cannot readily be attributable to specific activities, given limitations in data or accounting methods.
- Unlike targets based on individual performance measures, global efficiency measures can offer local policy makers the freedom to set their own priorities, and to seek out improvements along dimensions of performance where they believe gains are most readily secured.
- Measures of efficiency can be used to support other objectives, such as allocating finance or identifying the priority organisations for performance improvement.

Information about relative performance is often summarised in the form of a league table. This presentational format can be a powerful tool in its own right:

- League tables place system performance at the centre of the policy arena. For example, if countries had not been ranked in the World Health Report 2000, it is highly unlikely that the World Health Organisation (WHO) would have been able to generate the technical and policy debate that resulted (World Health Organization, 2000).
- League tables facilitate communication with ordinary citizens and promote accountability.
- The high profile secured by league tables can stimulate the search for better data and better analytic efforts.

In the past, the National Health Service (NHS) has sought to develop comprehensive efficiency measures, the labour productivity index and the purchaser efficiency index being recent examples. These indices assess ‘efficiency’ by dividing outputs by inputs or cost. When a range of outputs is produced, it becomes necessary to weight them in some way prior to aggregation. These weights indicate the relative value of a unit of each type of output. The labour productivity index and the purchaser efficiency index suffer problems common to all such indices. First, there will be trade-offs in the pursuit of different objectives, necessitating that they are weighted in some way. Usually, the selection of output weights will be controversial, particularly if there are sound arguments for allowing the importance of different outputs to vary among the organisations under consideration. Second, the index assumes that a simple relationship between outputs and inputs holds at all levels of operation. For instance, there is usually no recognition that larger organisations may be more (in)efficient than smaller ones simply by virtue of the scale of their activities. Third, there may be exogenous (or environmental) factors that inhibit or promote organisational attainment. Questions arise as to what these factors might be and how they should be taken into account.

Broadly speaking the policymaker’s concern is to construct an index of cost-effectiveness that for observation 0 takes the form:

$$P_0 = \frac{\sum_{r=1}^{s} y_{r0} w_r}{X_0}$$

where s is the number of outputs, $\boldsymbol{y_0}$ is the vector of outputs for observation 0, $\boldsymbol{w}$ is a vector of weights reflecting policy preferences, and X_0 is the expenditure of observation 0. This basic formulation will have to be modified slightly if there are additional exogenous inputs that influence performance. The variables contained in index may be manipulated in various ways (by taking logarithms or introducing higher order powers, for instance) but policy makers may be resistant to this if they feel it renders the index too opaque to be understood by those whose performance they wish to assess and influence. A league table of performance comprises a ranking of organisations according to this index.

The economics and operational research literature has sought to derive this index by developing two schools of analytic thought: econometric or *parametric* methods, such as Stochastic Frontier Analysis (SFA), that use multivariate regression models to explore why output or costs differ across organisations; and *non-parametric* methods, pre-eminently Data Envelopment Analysis (DEA), that attempt to measure efficiency by estimating the optimal level of output conditional upon the amount and mix of inputs.

Policy makers are showing increasing interest in these techniques. In 2000, the Public Services Productivity Panel produced a report in which the efficiency of the police service was analysed (Spottiswoode, 2000). The study recommended "the joint use of two of the most advanced relative efficiency measuring techniques – Stochastic Frontier Analysis and Data Envelopment Analysis" (page 4). Significant claims were made about their value, suggesting that the techniques would "provide a systematic, comprehensive measure of relative police efficiency ... and allow differentiated efficiency (performance) targets to be set for the police" (page 5).

This recommendation was well received in policy circles, with the Chief Secretary to the Treasury stating that the approach "clearly has wide potential application across all public services" (Spottiswoode, 2000). This endorsement implies that the techniques could form an important component of the strategy developed by the Public Services Productivity Panel in fulfilling its remit to advise on improving efficiency and productivity across government departments.

The UK is not alone in attempting to generate global measures of efficiency. One of the highest profile examples is a study conducted by the WHO. The WHO attempted to measure the overall performance and efficiency of national health systems and produced rankings of each country in terms of how well it was assessed to have promoted population health and other health system objectives (Evans, Tandon, Murray, & Lauer, 2001, World Health Organization, 2000). An econometric method was adopted to generate these rankings, and the methodological papers supporting the main results included a brief overview of the parametric and non-parametric techniques available for this form of analysis (Evans, Tandon, Murray, & Lauer, 2000, Tandon, Murray, Lauer, & Evans, 2000).

Given that these methods are gaining increasing acceptance in policy-making circles, it is important to understand their methodological underpinnings and to be able to evaluate their ability to provide an accurate assessment of overall organisational performance. The next section summarises the techniques before we turn to describing some of the main issues of contention in using these methods.

3 Brief overview of the techniques

This section outlines the assumptions underlying parametric and non-parametric techniques. Each has its own strengths and limitations. Moreover, within each empirical framework, a series of estimation decisions must be made, and there is no generally accepted methodology for guiding such decisions. This is not to imply that estimation decisions are arbitrary. The overarching criterion for guiding technical judgements should in principle be that the total expected costs of incorrect inference are minimized. Technical choices reflect trade-offs: by choosing to estimate a particular model in order to overcome one type of problem, a different type of problem might emerge. The analyst's task is to judge which technical choices balance these potential costs in an optimal fashion.

The difficulty in recommending a straightforward process toward efficiency measurement is not surprising in view of the nature of the phenomenon being studied. Inefficiency is inherently unobservable. This means that estimates of inefficiency have to be derived indirectly, after taking account of observable phenomena. In crude terms, this involves the following process:

- measuring observable phenomena (outputs, inputs, costs, prices ...);
- specifying some form of relationship between these phenomena;
- defining 'efficient' behaviour;
- calculating the difference between each unit's observed data and the maximum achievable as defined by the specified relationship;
- judging how much of the difference is attributable to inefficiency.

The main parametric and non-parametric methods are described briefly in this section. More detail on the theoretical underpinnings of the approaches appears elsewhere (Greene, 1993, Hollingsworth, Dawson, & Maniadakis, 1999).

3.1 The parametric approach

The parametric approach entails the following general process:

- Identify a *dependent variable* – either output or cost (y).
- Specify a set of *explanatory variables* ($\boldsymbol{x}$) that are thought to explain or predict differences in output or cost.
- Interpret residual differences between observed and predicted output or cost as arising from either measurement error or inefficiency (e).

The dependent and independent variables are related by specifying an econometric model of the general form:

$$y_i = a + bx_i + e_i \qquad (1)$$

where y indicates either output or cost, $\boldsymbol{x}$ is a vector of explanatory variables, and the residual e represents the deviation between the observed data and the relationship predicted by the independent variables in the model.

A number of decisions must be made in order to estimate a parametric model of this form, such as whether to estimate a production or cost function, choice of functional form, which independent variables to include, whether to work in natural units or logarithms, whether to work in total or average costs, and how to interpret the residual. All such choices will be influenced by the scope and nature of data availability. The treatment of the residual is a distinguishing feature of the technique, making these models a class apart from usual econometric methods. In econometrics generally the residual is not afforded special attention, whereas in efficiency analyses it is often deemed the only parameter of interest. We shall return to the implications of this shortly, but first the reason for the interest in the residual is explained.

In early econometric applications of efficiency analysis, the entire residual was interpreted as being due to inefficiency. Under this so-called Corrected Ordinary Least Squares (COLS) approach, in the case of a production function, the organisation with the highest residual value is defined as being fully efficient. Its output is both higher than that predicted by the model and higher than that for any other observation, holding constant the variables in the model. This implies that the efficiency frontier is located by shifting the regression line so that it passes through this fully efficient observation. This is illustrated in figure 1, where observation C is efficient. The inefficiency of the remaining observations can be measured by their vertical distance from this frontier, as shown for observation A.

The COLS approach fails to recognise that the residual may incorporate factors other than inefficiency, such as measurement error or omitted model variables. If this is the case (as is likely), it is inappropriate to interpret the entire residual as representing inefficiency. Stochastic Frontier Analysis was therefore developed as a means to separate the residual into two components: inefficiency and all other sources of model error. This is achieved by introducing the assumption that the efficiency component of the residual is distributed asymmetrically whilst the remaining error is normally distributed in the usual symmetric fashion. Figure 2 provides a simple illustration of the technique. The stochastic frontier has two notable features. First, it does not correspond to the 'line of best fit' through the observations that would be produced by a simple linear regression model. Second, the frontier does not (necessarily) pass through the observation that produces the maximum level of output conditional upon input (observation C). This is because the frontier is estimated after recognising that some of the difference between observed output and the level of output predicted by the explanatory variables may be due to measurement error. In the figure below, observation C lies above the estimated frontier. The distance of this point from the frontier is attributable to measurement error. For observations lying below the frontier, the distance comprises both measurement error and inefficiency, as illustrated for observation A.

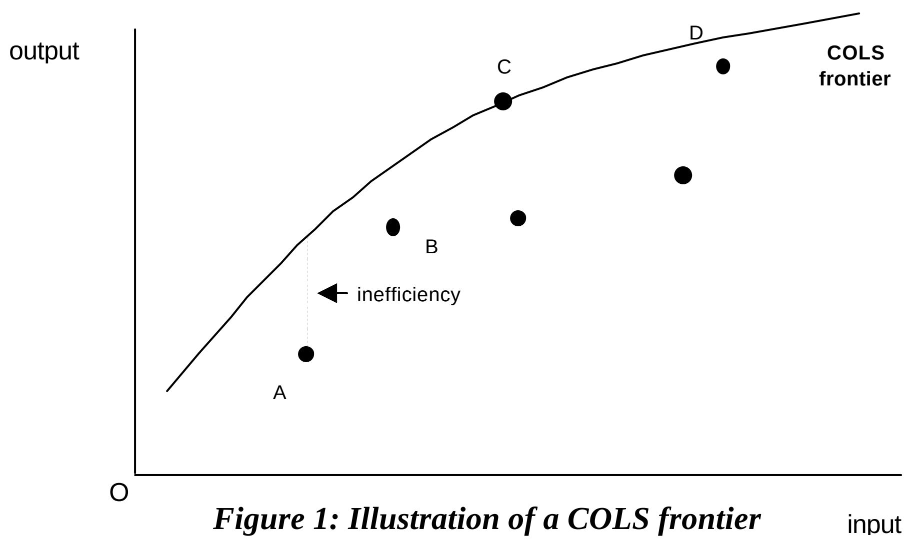

Figure 1: Illustration of a COLS frontier

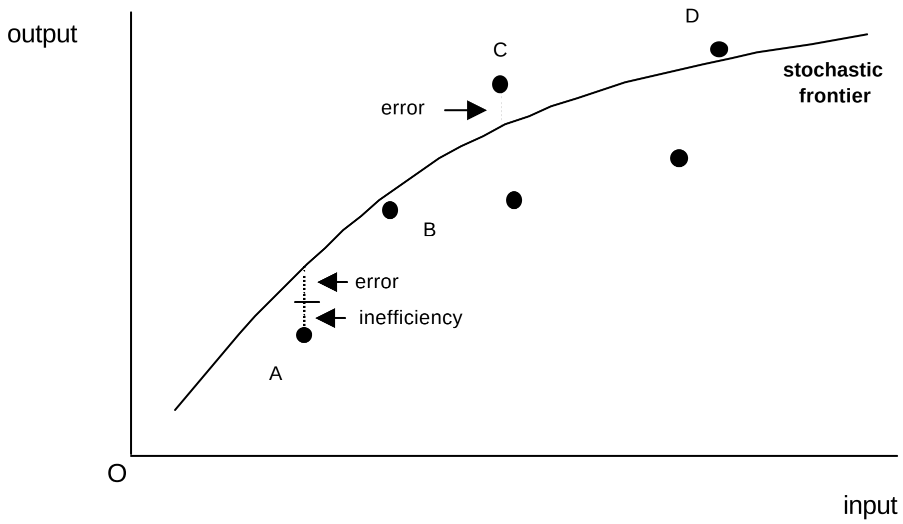

Figure 2: Illustration of a stochastic frontier

3.2 The non-parametric approach

The parametric approach infers the efficient frontier from the behaviour of all observed organisations. In contrast, under data envelopment analysis the location and shape of the efficiency frontier is determined only by extreme observations. DEA is based on the simple notion that an organisation that employs less input than another to produce the same amount of output can be considered more efficient. The efficiency frontier is constructed of linear segments that join up those observations with the highest ratios of output to input. The resulting frontier thus 'envelopes' all the observations. In figure 3, observations A, B, C and D are considered efficient, given the scale of their operations. The inefficiency of observation E is indicated by either its vertical or horizontal distance from the frontier – it uses more input to produce a similar level of output to observation B and, despite employing a similar amount of input to observation C, it produces considerably less output.

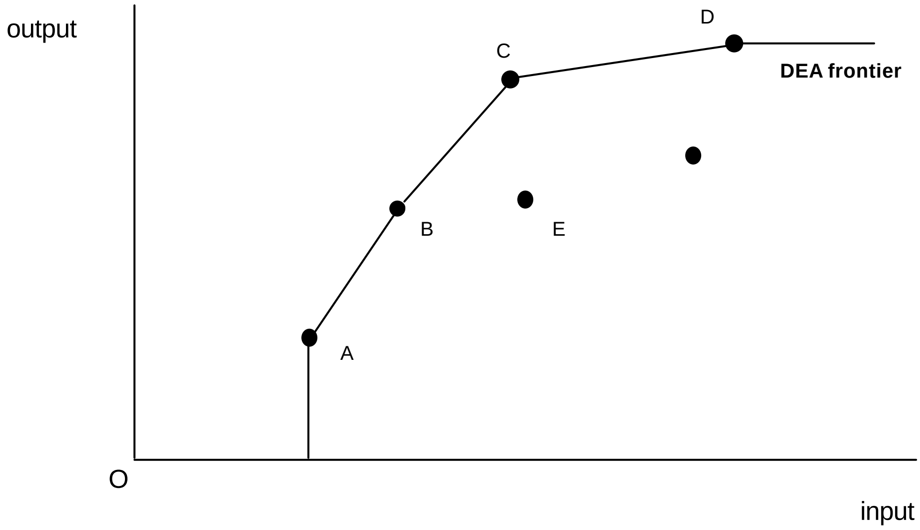

Figure 3: Illustration of DEA frontier

Assuming just one input X, the DEA formulation of the production process seeks for each organisation 0 to find a set of weights $\boldsymbol{u^0}$ and an efficiency score q_o so as to:

$$\text{maximise} \quad q_0 = \frac{\sum_{r=1}^{s} y_{r0} u_r^0}{X_0}$$

$$\text{subject to} \quad \frac{\sum_{r=1}^{s} y_{rj} u_r^0}{X_j} \leq 1 \quad j = 1,...,n$$

where $\boldsymbol{u^0}$ is the set of weights that maximizes the apparent performance of organisation 0, subject to requiring that the efficiency score for all organisations is no greater than 1 using that set of weights.

Note that in general a different set of weights $\boldsymbol{u^j}$ will be computed for each observation j. The DEA score q_o yields a performance index for unit 0 subject to a specific set of weights $\boldsymbol{u^0}$ but, in general, each member of the array of scores $\{q_j\}$ will be constructed using a different set of weights. This raises the question of whether it is appropriate to rank DEA efficiency scores in a conventional league table format.

4 Discussion of the techniques

In this section we discuss four of the most important issues that arise when seeking to use productivity models in health and health care: the weights used to indicate the values of different outputs; how the efficiency models are constructed; the treatment of environmental influences on performance; and dynamic aspects of productivity.

4.1 Output weights

There are important questions relating to the objectives encompassed by the index of cost-effectiveness. Is it legitimate for the central policy maker to attach a uniform set of objectives to all organisations? If so, is it further legitimate to apply a uniform set of weights? If so, how should they be chosen? If not, what is the extent of legitimate variation, and who should choose? These are fundamental issues, the answers to which determine whether or not a 'league table' approach is warranted. In our view, observations can be ranked only if the policy maker may legitimately (a) set objectives and (b) attach weights to those objectives.

The set of weights $\boldsymbol{w}$ is central to the operation of any index of performance. If estimating a cost function, the weights are derived from the parameter estimates. Using a linear model, the weight $\boldsymbol{b}_r$ attached to output r indicates the value of an additional unit of that output, which remains constant for all levels of attainment of y_r. If a logarithmic model is used, $\boldsymbol{b}_r$ indicates the percentage increase in composite attainment implied by a one percent increase in y_r. If econometric methods are used, the estimated magnitude of the weight will be the value implicit in the sample mean cost of producing an additional unit of output r. In conventional DEA the analogous weights, u_r, are allowed to vary freely among observations.

In practice, very few studies have paid serious attention to the validity of weights. The use of simple econometric methods implies a belief that the expenditure patterns of organisations (on average) reflect the values placed by society on the outputs. In DEA there has been some attention to the notion of weight restrictions, but the efforts to date have been poorly informed by economic theory, and mainly confined to technical considerations. There do exist more theoretically sound approaches to inferring weights (such as economic studies of willingness to pay or conjoint analysis), but rarely have these been used in practical applications.

Moreover, the weights and the scale of measurement are inextricably linked, but little attention has been paid to this interaction. The weights $\boldsymbol{w}$ should be valid at all levels of attainment. Given the policy imperative to use simple models of performance, this implies a need for great care in the scale of measurement for each input and output. Many performance indices use measures of output that have been transformed (for example, z-scores, logarithmic, or linear transformations). However, it is unusual to see any attention paid to whether the weights used continue to be valid at all levels of attainment after such transformation, as required.

Consider the economist's traditional production possibility frontier FF for two outputs, reflecting two societal objectives, as shown in Figure 4. We assume constant returns to scale. Two sets of preferences are illustrated by indifference curves I_1I_1 and I_2I_2, giving rise to different preferred points of production. The slopes of these curves at the points of tangency with FF reflect the relative valuations of the two system outcomes. In this case, individual 1 places a higher relative valuation on outcome 1 than individual 2. In general, there will be no agreement on what constitutes the preferred mix of outputs.

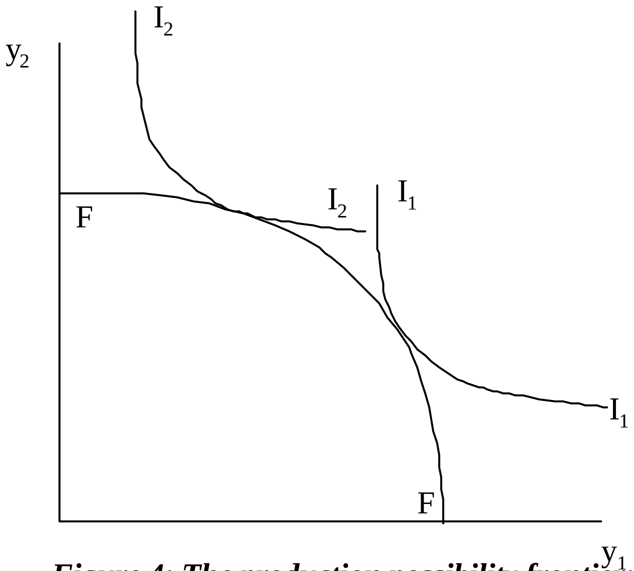

Figure 4: The production possibility frontier: different preferences lead to different weights

The use of a linear performance index suggests that resolution of the trade-off problem should be guided by maximizing a linear function of the two outcome measures, which are combined into a single composite indicator. The parallel lines in Figure 5 indicate different values of a chosen composite indicator, with scores increasing towards the top right hand corner. Choice of the point P^* on the possibility frontier would be optimal in this example, giving a composite score indicated by the line C_1C_1. Given the weights used in the composite indicator, choice of any other point on the frontier would be considered inferior (allocatively inefficient).

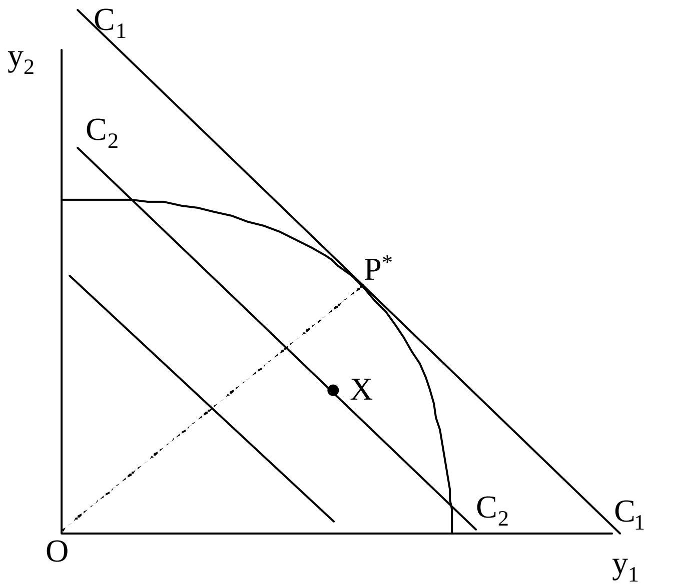

Figure 5: Composite scores indicated by the lines C_1C_1 and C_2C_2

The implication of imposing a set of weights is that few organisations will be precisely on the possibility frontier. Rather, each will exhibit some level of inefficiency, which leads to observed outcomes lying within the area indicated by the efficient frontier (technical inefficiency). In the context of the diagram, the point X indicates a realized level of performance in one organisation. According to the composite indicator, this secures a level of system efficiency indicated by the line C_2C_2, reflecting the fact that (a) the chosen mix of outputs diverges from the 'optimal' and (b) performance lies within the frontier. The measure of organisational efficiency can be represented by the ratio of the composite scores indicated by lines C_2C_2 and C_1C_1, the extent to which performance falls short of the maximum attainable and desired. It is the product of technical and allocative efficiency. This argument is readily extended to *n* outputs.

If we are unable to apply a uniform set of weights, nevertheless there may be circumstances in which all will agree that some organisations perform better than others. Figure 6 illustrates five organisations with identical expenditure levels and environmental circumstances. Under most assumptions about preferences, organisation A is unambiguously inferior to organisation D in the sense of being technically inefficient. Furthermore, organisation B is inferior to a linear combination of organisations D and E, represented by the point B^*. However, the ranking of the organisations C, D and E lying on the observed frontier depends on the relative weights we choose to apply to outcomes 1 and 2. This is the analytical principle underlying DEA. However, we cannot rank these organisations without introducing a composite indicator that reflects preferences for outcome 1 relative to outcome 2.

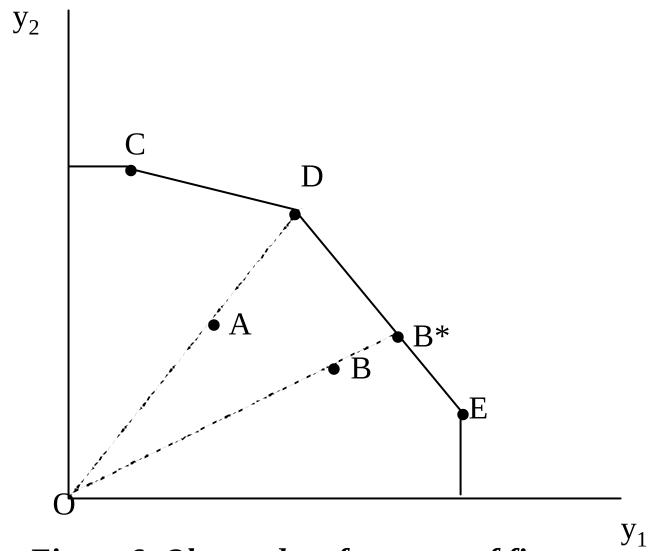

Figure 6: Observed performance of five systems with identical expenditure and environment

4.2 Modelling the production process

Having decided upon what objectives are to be considered, the next problem concerns how to model the process by which these may be achieved and/or the constraints that limit levels of attainment.

The research interest in productivity models is predominantly in the structure and determinants of the production process rather than specific efficiency estimates for individual organisations. Countless research questions present themselves. For example: What is the marginal productivity of a factor of production? How do returns to scale vary? What influence do external environmental factors have on productivity? These are all important questions with potentially important policy implications. However, they all fit into the traditional empirical research model in that they seek to identify aggregate (or sample average) patterns within the data. That is, although individual inefficiency might be modeled explicitly (for example using a one-sided error term), such modeling is usually a means to the end of securing a more satisfactory aggregate model with which to address the research questions.

In contrast, the managerial or policy interest is in the estimate of efficiency for individual organisations. This estimate is derived from the residual or fixed effect, and the model parameters are no longer the main interest. This switch of attention turns the econometric model on its head. Our contention is that this may require a fundamental rethink in modeling methodology.

Traditional econometric methodology seeks to develop an empirical model that satisfies particular acceptability criteria, such as consistency (as the sample size increases, does the estimate of interest converge to its "true" value"?); unbiasedness (is the expected error in the estimate zero?); efficiency (is the sampling variance of the estimate as small as possible?); robustness (is the estimate robust to potential model misspecification, missing information and measurement error?); and parsimony (is the model as simple as possible?). Although heuristic rules of thumb are frequently used by analysts (such as the 95% significance criterion), the implications of technical choices for model estimation are generally well understood, so that an informed observer can understand the degree of certainty with which inferences can be made.

However, there is no guarantee that an econometric model that satisfies traditional modeling criteria is necessarily fit for the purpose of inferring efficiencies of individual organisations. To take just a simple example, there might exist a small number of ambulance authorities that suffer a cost disadvantage in their emergency function. In developing a cost function, the analyst might acknowledge this possibility and test a measure of rurality as a potential independent variable. This variable may not pass the 95% significance hurdle, and its exclusion leads to a more parsimonious model that passes misspecification tests. The analyst excludes it from the preferred model on the grounds that it is immaterial. However, of course, it may be highly material for the small number of ambulance authorities whose efficiency estimates (residuals) are adversely affected by its exclusion. The model is therefore fit for its research purpose, but not for its managerial purpose.

Conversely, one could pay no attention to the parsimony criterion and indiscriminately include all potential explanatory variables in the productivity model. In the extreme, this might result in modeling the productivity of all observations without error, leading to conclusion that all are equally efficient. If, in reality, there is some heterogeneity in efficiency, the inability to detect it arises because some of the explanatory variables are correlated with efficiency. One therefore needs a very clear idea of the production process and the constraints upon that process if one is to model individual efficiency satisfactorily.

As with all modeling, the ideal is that technical choices should be informed by the costs of incorrect inference. In the traditional model-building methodology, in the Neyman and Pearson (1931) tradition, notions such as significance level and power are used to give an indication of the probability that inferences may be mistaken. In crude terms, technical choices should seek to balance the costs of type I and type II errors.

In all likelihood, every organisation exhibits a level of inefficiency with respect to a true production frontier that is unobservable. The managerial concern is in the extent to which the chosen model misrepresents this true efficiency. An underestimate of individual efficiency may result in a number of mistaken managerial actions, such as setting financial penalties, replacement of local management, demanding infeasible improvement targets, or closure of the operation. An overestimate of efficiency may result in complacency or mistaken designation of an organisation as a beacon of excellence. Errors of either sign can arise from model misspecification (omitted variables, functional form) or measurement errors.

4.3 Environmental constraints

Numerous classes of factors influence observed levels of attainment. These include:

- differences in the characteristics of citizens being served;
- the external environment – for example, geography, climate, culture;
- the activities of other related agencies;
- the quality of resources being used;
- different accounting treatments;
- data errors;
- random (or idiosyncratic) fluctuation;
- different priorities;
- differences in efficiency.

In the short run, many of these factors are outside the control of the organisations under scrutiny. We call these ‘environmental’ variables. In the longer term a broader set of factors are potentially under the control of the organisations, but the extent and nature of this control will vary depending on the nature of the analysis.

In whatever way that the uncontrollable ‘environment’ is defined, it is usually the case that some organisations operate in more adverse environments than others, in the sense that external circumstances make achievement of a given level of attainment more difficult. This means that – for a given level of expenditure – the production possibility frontiers of different organisations will not be identical. The frontiers for organisations operating in difficult environments will lie inside those of more favourably endowed organisations.

There is an active debate about how to incorporate environmental factors into a DEA analysis. One option is to include an environmental variable as one of the inputs in the production model. This means that organisations will be compared only with organisations operating in identical or more adverse environments. Those operating in the most adverse environments will automatically be deemed 100% efficient. Another possibility is to incorporate environmental variables only in a second stage analysis, which seeks to explain efficiency scores as a function of environmental variables, usually be applying econometric techniques. This is problematic given that the dependent variable (the efficiency scores) will comprise a set of serially correlated values (Simar & Wilson, 2002). The essential point to note is that there is no generally accepted methodology for how to account for environmental variables in DEA models or how to test whether an environmental variable is a 'significant' influence on production possibilities.

Under the econometric paradigm, when the objective of the exercise is to make inferences about relative efficiency, a necessary condition is that all variables included as regressors are exogenous to organisational control (Giuffrida, Gravelle, & Sutton, 2000). Again this contrasts with the traditional approach to econometric model building, where the aim is to select a set of explanatory variables that best explain variation in the dependent variable.

Nevertheless, efficiency models rely on traditional econometric selection devices to test whether an environmental variable should be included in a SFA model. This may not be appropriate. Recall that SFA decomposes unexplained variation from predicted output into two parts: random error and inefficiency. Loosely speaking, symmetric (two-sided) error is attributed to randomness, one-sided error to inefficiency. Suppose therefore that we wish to test an environmental variable for inclusion in a SFA model. It will be a candidate for inclusion if it 'explains' a material proportion of the unexplained error, whether one-sided or two-sided, and therefore exhibits what is conventionally termed a statistically significant model coefficient.

However, the attribution of a statistically significant effect to the explanatory variable may be for one or both of the following reasons:

- it explains some of the random error (that is the original model suffered from omitted variable or functional form misspecification);
- it explains some of the inefficiency error (that is, the variable is correlated with the original estimates of inefficiency).

In the latter case, the new variable explains away some of the inefficiency identified in the original model. This may be because (a) it represents some unavoidable hindrance to reaching the estimated frontier, such as an adverse operating environment, or (b) it represents some potentially controllable characteristic of inefficient organisations, such as a poor geographical configuration of capital resources.

In principle, we should want to include type (a) variables in the model, so that the estimated frontier reflects the uncontrollable environmental circumstances in which an organisation must operate. In contrast, we would want to exclude type (b) variables. Again, unless we have a clear idea of the production process under scrutiny, we have little guidance on choices.

An obvious response to uncertainty about how best to represent the underlying set of production possibilities or technological constraints is to conduct sensitivity analysis. In DEA, this may involve changes to the scaling assumptions and bootstrapping estimates to assess statistical significance. Econometric techniques allow different functional forms, different distributions of inefficiency, and the calculation of confidence intervals around inefficiency estimates.

The extent to which results are robust to these choices depends on the complexity of the underlying production process. In industries with a relatively simple production technology, it can be expected that results are not highly sensitive to defensible (ie non-arbitrary) technical choices. For instance, estimates of the efficiency of companies providing water and sewerage services in the UK appear robust in the face of sensitivity analysis (Office of Water Trading, 1999). In contrast, efficiency estimates derived from different models the UK hospital sector rarely yield definitive or consistent conclusions (Jacobs, 2001). Quite modest changes in the choice of analytic technique and model specification lead to major changes in inference about efficiency. This stems from the fundamental problem in specifying how hospital care is produced. As Harris argues "business as usual in hospital is ... a continuous sequence of potential crises ... [with] hospital care involving a complicated sequence of adaptive responses in the face of uncertainty ... In contrast to the standardised assembly line production process, each patient receives customised attention" (Harris, 1977). This implies that generalised models of production will have difficulty in capturing the idiosyncratic nature of providing individualised packages of treatment. In contexts where the production process is inaccurately represented or production constraints are poorly understood, efficiency estimates will be more sensitive to analytical choices.

It should be recognised that, in some situations, including environmental factors may be unnecessary. It may be possible to simplify the 'environmental' problem when organisations (such as health authorities or primary care trusts) have already been compensated financially for environmental circumstances through a funding formula. A funding formula seeks to enable organisations to deliver some 'standard' level of service, given environmental factors. So, if the funding formula is doing its job properly, there is no need to incorporate such factors into the productivity model. Indeed, all that is needed may be to examine the extent to which the standards have been secured. In short, one may need to examine only effectiveness, and not incorporate inputs (either resources or environment) into the model at all.

4.4 Dynamic effects

One of the most problematic issues in productivity analysis is the treatment of dynamic effects. Dynamic aspects of performance are particularly important in many areas of health care, where current outcomes are highly dependent on past inputs and efficiency, and current inputs and efficiency are highly influential on future outcomes. An immediate consequence of this is that – unless treated with great care - cross-sectional analysis is likely to be inappropriate. However, a truly dynamic model may make infeasible demands on data and econometric methodology.

Suppose, for example, that performance P_t in the current period t is dependent on some weighted sum of effort $p(e_t)$ and stochastic shocks in the current and previous periods, as follows:

$$P_t = \sum_{i=0}^{t} a^{t-i}\{p(e_i) + e_i\}$$

where $\alpha < 1$ reflects the dissipating influence of previous effort on future performance. An alternative way of writing the evolving level of performance achieved is as:

$$P_t = aP_{t-1} + p(e_t) + e_t$$

That is, current performance is the consequence of current and previous managerial actions, which can be thought of as a weighted sum of past performance and current effort.

A further complication is that, to some extent, current effort may be directed towards future attainment. So any cross section of contemporary attainment will capture the outcome of historical effort and historical accident. And any cross section of contemporary resource consumption will contain an element of investment for *future* attainment. If, further, current effort contains an element of inefficiency (which may change over time), it becomes far from clear what a productivity model based on cross-sectional or limited panel data is seeking to capture.

In practice, of course, the dynamic production process is likely to be more complex even than this stylized representation suggests. However, it highlights some of the important dynamic considerations that are likely to impinge on observed current system behaviour. This analysis also indicates why conventional panel data techniques are likely to be inappropriate. The panel for a particular observation is not a sequence of random draws, but is rather a subtle time series. Econometric modelling and estimation of such processes is likely to be a challenging endeavour, even given adequate data.

5 Conclusions

Research efforts in productivity analysis have burgeoned in recent decades. Much of this research effort is to be applauded, and future endeavours should be encouraged. However, our intention in this paper has been to point out the poor understanding of the role that productivity analysis might have for policy purposes. In our view, both researchers and policy makers should be seeking to improve this understanding.

Is it legitimate for policy makers to seek to develop global measures of organisational cost-effectiveness? The answer to this question must depend on the institutional framework within which the analysis is being undertaken. For example, the UK National Health Service is a unitary organisation that seeks to secure uniform standards across the nation. In this context, it is reasonable to measure comparable institutions on a consistent basis. The position is less clear cut if the organisations under scrutiny have a degree of autonomy as to values (for example in a system of local governments), although even here it may be legitimate for a central authority to seek to measure departures from some set of national priorities.

The next question is to ask is why should policy makers have any interest in global measures of organisational cost-effectiveness? In may respects it may be enough to examine specific organisational functions (for example, the length of stay following hip replacement) rather than develop a summary measure of total organisational efficiency. We have indicated some reasons why policy makers may feel that such summary measures are useful, but few are directed towards helping managers improve specific aspects of performance. Much more detailed benchmarking data are needed for that managerial purpose. Rather, the immediate purpose of summary measures is often to heighten awareness of performance issues, to create league tables, and to offer information that may act as a spur for seeking out improved performance.

It is not necessary to use analytic techniques to develop indices of cost-effectiveness (Stone, 2002). The main pre-requisites for developing such an index are (a) the choice of outputs to be measured (b) choice of a set of weights reflecting the values attached to each output and (c) good measurement of the components of the index. In health care, rarely will there be consensus about (a) and (b), which are essentially political choices. The use of productivity models merely imposes a set of assumptions about output weights, rather than offering a solution to the political problem. To this end, alternative approaches, such as trade-off surveys or conjoint analysis, could be more useful to policy makers than productivity models. There will always be a need for a dialogue between analysts and policy makers in order to ensure that models deployed reflect policy priorities.

One area in which productivity models do appear to offer some help is in adjusting cost-effectiveness scores for environmental circumstances. However, we have sought to show that, in circumstances where it is unclear how such factors enter the production function, the treatment of environmental factors is highly problematic.

Dynamic aspects of health care can be very important, especially in the public health domain. The dynamic features of the system can be considered as a special type of environmental variable, indicating the historical inheritance or endowment of previous organisational efforts within which the organisation must operate. An important output of the organisation will then be the endowment it leaves for future management, as well as the current health outputs.

We have not dwelt on the profound measurement difficulties that often severely constrain the choice of productivity model (Newhouse, 1994). Yet these often pose some of the central difficulties in interpreting model results. For example, analysts are often constrained to working with cross-sections or short panels of data. Many aspects of output, particularly in the quality domain, are unmeasured or poorly measured. More generally, there are great uncertainties in the measurement of many important aspects of health organisations.

Finally we have sought to emphasize that the institutional rankings emerging from productivity models are in general highly volatile, depending on technical choices about which the analyst has little methodological guidance. There is a clear and urgent need for econometricians to reconsider model-building methodology when the interest is in the residuals rather than parameter estimates.

In the light of our discussion, we make the following recommendations to anyone seeking to use productivity models for policy purposes:

- Pay careful attention to the purpose of analysis and to how results are to be used;
- Seek to develop a realistic and theoretically coherent model of production;
- Ensure that an econometric methodological process is followed;
- Leave a careful audit trail of technical choices;
- Consider a broader portfolio of approaches towards developing indices of cost-effectiveness;
- Seek policy guidance on judgements that are properly political in nature;
- Ensure that policy makers do not rely solely on productivity models as a basis for making judgements about organisational efficiency.

References

Evans, D.B., A.J. Tandon, C.J.L. Murray, & J.A. Lauer. 2001. Comparative efficiency of national health systems: cross national econometric analysis. *British Medical Journal*, 323: 307-10.

Evans, D.B., A.J. Tandon, C.J.L. Murray, & J.A. Lauer. 2000. *The comparative efficiency of national health systems in producing health: an analysis of 191 countries*. Geneva: WHO.

Giuffrida, A., H. Gravelle, & M. Sutton. 2000. Efficiency and administrative costs in primary care. *Journal of Health Economics*, 19: 983-1006.

Greene, W.H. 1993. The econometric approach to efficiency analysis.In Fried, H.O., C.A.K. Lovell, & S.S. Schmidt, editors, *The measurement of productive efficiency: Techniques and applications*. New York: Oxford University Press.

Harris, J.E. 1977. The internal organisation of hospitals: some economic implications. *Bell Journal of Economics*: 467-82.

Hollingsworth, B., P.J. Dawson, & N. Maniadakis. 1999. Efficiency measurement of health care: a review of non-parametric methods and applications. *Health Care Management Science,* 2: 161-72.

Jacobs, R. 2001. Alternative methods to examine hospital efficiency: data envelopment analysis and stochastic frontier analysis. *Health Care Management Science*, 2: 103-15.

Newhouse, J.P. 1994. Frontier analysis: how useful a tool for health economics? *Journal of Health Economics*, 13: 317-22.

Office of Water Trading. 1999. *Future Water and Sewerage Charges 2000-05: draft determination*. London: OFWAT.

Simar, L. & P.W. Wilson. 2002. Estimation and inference in two-stage semi-parametric models of production processes. Louvain: Institut de Statistique, Universite Catholique de Louvain.

Spottiswoode, C. 2000. Improving police performance: a new approach to measuring police efficiency. London: Public Services Productivity Panel, HM Treasury.

Stone, M. 2002. How not to measure the efficiency of public services (and how one might). *Journal of the Royal Statistical Society,* Series A, 165(3): 405-34.

Tandon, A.J., C.J.L. Murray, J.A. Lauer, & D.B. Evans. 2000. *Measuring overall health system performance for 191 countries*. Geneva: WHO.

World Health Organization. 2000. *World Health Report 2000*. Geneva: WHO.