Producing health services efficiently: a review of measurement tools and empirical applications

Damian Walker
London School of Hygiene and Tropical Medicine

Rianna Lisa Mohammed
Judd Institute of Management Studies
University of Cambridge

December 2004

Published by Oxford Policy Institute

ISBN 978-0-9551123-0-0

Abstract

Healthcare costs in most developed economies have grown dramatically over the last few decades, and it is widely believed that the inefficiency of healthcare institutions, has, at least in part, contributed to this phenomenon. In response, there has emerged, in recent years, a growing body of literature on the efficiency of health care services in industrialised countries, particularly in the US. Unfortunately, there has not been a similar focus on efficiency in the production of health care services in less-developed economies. This is particularly disappointing given the developing world's greater scarcity of financial resources with the inefficient use of scarce resources exacting a much higher penalty in terms of foregone health benefits. Productivity and efficiency improvements are thus critical, given resource constraints faced by the public health sector in many developing countries. In short, improving the efficiency of health services in developed and developing countries should be a major goal of public, private and non-profit providers alike. Knowledge of the levels and determinants of health services efficiency can help policy-makers and health care managers take measures aimed at curtailing costs while maintaining acceptable levels of quality and access. However, there are, methodological problems that make the measurement of both health services productivity and efficiency difficult. Against this backdrop, the Oxford Policy Institute, with sponsorship from the Economic and Social Research Council, organised a series of seminars between January – May 2004. The overall aim was to lay the foundation for an international comparative research portfolio designed to identify the key ways in which health services can be delivered more efficiently in different settings and to disseminate that knowledge widely. The objectives of the series were to review : techniques for measuring productivity and efficiency; the literature on the productivity and efficiency in OECD, transitional and low-income countries; the effects on productivity of changes in skill mix and incentives; and the effects of managerial and technological innovations on productivity. The purpose of this monograph is to summarise some of the main findings from the seminar series. In particular, the monograph provides: definitions of efficiency-related concepts; a summary of the alternative approaches to efficiency measurement, including a discussion of broad-based methodological issues in the measurement of efficiency; a selective review of intra- and inter-country efficiency analyses, including evidence on the variation of efficiency over-time and across a group of countries; and finally, some thoughts on the way forward for researchers and policy-makers are considered.

Acknowledgements

Damian Walker is a member of the Health Economics and Financing Programme (HEFP), which is supported by programme funds from the Department for International Development, UK (DFID).

Rianna Lisa Mohammed is a PhD student at the University of Cambridge, UK. Her research is focused on the monitoring and evaluation of HIV/AIDS in developing countries.

The Economic and Social Research Council funded the seminar series on which this monograph is based (for further details go to www.opi.org.uk).

The seminar speakers were: Karen Bloor, Roy Carr-Hill, Hugh Gravelle, Alan Maynard & Andrew Street (University of York); Bruce Hollingsworth (Monash University); Alistair McGuire & Maria Raikou (LSE Health and Social Care, LSE); Nicolai Mai (Office of National Statistics); Ravi P. Rannan-Eliya & Aparnaa Somanathan (Institute of Policy Studies, Sri Lanka); D. Varatharajan (Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Thiruvananthapuram, Kerala, India); and Carol Propper (CMPO, University of Bristol).

Contents

Producing health services efficiently: a review of measurement tools and empirical applications

Introduction

Healthcare costs in most developed economies have grown dramatically over the last few decades, and it is widely believed that the inefficiency of healthcare institutions, has, at least in part, contributed to this phenomenon (Worthington 2004). In response, there has emerged, in recent years, a growing body of literature on the efficiency of health care services in industrialised countries, particularly in the US (Hollingsworth 2004).

Unfortunately, there has not been a similar focus on efficiency in the production of health care services in less-developed economies. This is particularly disappointing given the developing world's greater scarcity of financial resources with the inefficient use of scarce resources exacting a much higher penalty in terms of foregone health benefits. Productivity and efficiency improvements are thus critical, given resource constraints faced by the public health sector in many developing countries.

In short, improving the efficiency of health services in developed and developing countries should be a major goal of public, private and non-profit providers alike. Knowledge of the levels and determinants of health services efficiency can help policy-makers and health care managers take measures aimed at curtailing costs while maintaining acceptable levels of quality and access. However, there are, methodological problems that make the measurement of both health services productivity and efficiency difficult.

Against this backdrop, the Oxford Policy Institute, with sponsorship from the Economic and Social Research Council, organised a series of seminars between January – May 2004. The overall aim being to lay the foundation for an international comparative research portfolio designed to identify the key ways in which health services can be delivered more efficiently in different settings and to disseminate that knowledge widely. The objectives of the series were to review: techniques for measuring productivity and efficiency; the literature on the productivity and efficiency in OECD, transitional and low-income countries; the effects on productivity of changes in skill mix and incentives; and the effects of managerial and technological innovations on productivity.

The purpose of this monograph is to summarise some of the main findings from the seminar series[1]. The first section will provide definitions of efficiency-related concepts. This will be followed by a discussion on alternative approaches to efficiency measurement, including a discussion of broad-based methodological issues in the measurement of efficiency. The next section will provide a selective review of intra- and inter-country efficiency analyses, and will also present some evidence on the variation of efficiency over-time and across a group of countries. Finally, the way forward for researchers and policy-makers will be considered.

The measurement of health services productivity and efficiency

The two important concepts to consider when analysing the efficiency of a decision-making unit (DMU)[2] are *technical* and *allocative* efficiency.

Technical efficiency

In order to measure efficiency, a norm must be specified. The norm set for measuring technical efficiency is that the minimum amount of resources should be used to produce a given level of output or, alternatively, the maximum amount of output should be produced for a given level of resource use. If more resources than necessary are used to produce a given amount of output, this implies a waste of resources and therefore inefficiency. Thus, the difference in the amount of output that could have been produced from a given amount of resources and the amount of output that was actually produced can be used as a measure of technical inefficiency. Technical inefficiency is therefore a matter of degree depending upon how much unnecessary resources have been used. Central to the measurement of technical efficiency is the notion of the *production possibilities frontier* or *isoquant*.

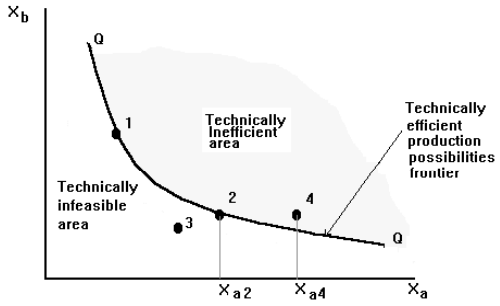
This is illustrated in Figure 1, for a simple production process that uses only two inputs, X_a and X_b (for example, these inputs could be doctor and nursehours worked, or a doctor's hours and drugs, etc.). Any point along the production possibilities frontier QQ represents a technically efficient way of combining various quantities of inputs X_a and X_b to produce the same amount of output Q. For example, while points 1 and 2 differ in the combination of X_a and X_b (production at 1 is more intensive in X_b than at 2), both permit production of the same quantity Q. Points 1 and 2, like all other points on the frontier QQ are technically efficient because it is not possible to produce Q with smaller quantities of either X_a or X_b, as depicted

1. Detailed notes on each seminar are available on www.opi.org.uk.
2. The term used to describe a productive entity in instances when the term 'firm' may not be entirely appropriate, e.g. when comparing the performance of public vaccination sites, the units are really *parts* of a firm rather than firms themselves.

by the line (there is no room for further gain in technical efficiency). Point 3, like all points to the left of the production possibilities frontier, is infeasible: any reduction in the amounts of X_a and X_b from the amounts represented by the frontier necessarily translates into a reduction in Q. In contrast, point 4, like all points to the right of the production possibilities frontier, constitutes a technically inefficient way of producing Q: technical efficiency can be improved by moving production from 4 to 2, thereby reducing the amount of X_a from X_{a4} to X_{a2}. In effect, one procedure is considered more technically efficient than another, if it either produces the same quantity of output using fewer inputs, or produces a greater quantity of outputs using the same resources.

The production possibilities frontier represents all the possible combinations of inputs, which permit production of the same quantity of health care output. It is important to note, that it is assumed here that technical quality of care also remains constant along the production possibilities frontier. Thus, not only does any combination of inputs X_a and X_b along the curve permit production of quantity Q of medical care output, but also, any such combination delivers medical care of constant technical quality, i.e., with the same effect on patients' health status.

Figure 1: Technically efficient production possibilities frontier or isoquant



Allocative efficiency

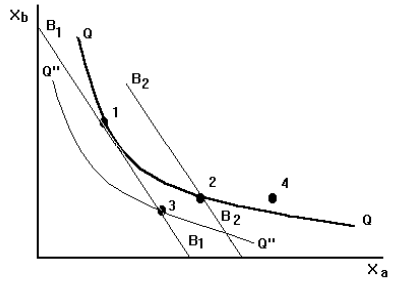
Similarly, an allocatively efficient DMU will combine these inputs in a cost-minimising manner to produce a given level of output, with price ratios being the norms for judging allocative efficiency[3]. With factor input prices given, resources used in production should be combined so as to reflect the corresponding ratio of different factor input prices. A mix of resource use

3. However, using prices as the criteria for measuring economic efficiency is based on the assumption that firms have no influence on the price. Rather, prices are determined in the market as the outcome of competitive bidding between a large number of consumers and firms – clearly this may not apply in the health sector.

that deviates from the corresponding ratio of given factor input prices is taken as a measure of allocative inefficiency. Any deviation in the mix of resource use from observed price ratios is measurable, and hence, allocative inefficiency becomes a matter of degree, just like technical inefficiency. Although there may be many technically efficient alternatives to produce a given quantity Q, there is generally only *one* allocatively efficient way of doing so.

Figure 2 helps to illustrate the fundamental difference and relationship between technical and allocative efficiency. Suppose that the unit prices of inputs X_a and X_b are W_a and W_b, respectively. If a health facility is allocated a budget B_1, then B_1 represents the facility's budget constraint. The constraint is given by the equation: $B_1 = X_a * W_a + X_b * W_b$. Any point along the budget constraint line, such as points 1 and 3, consumes the whole budget B_1. However, point 1 is preferable to 3 because at 1 quantity Q is produced, whereas at point 3 the smaller quantity Q" is produced. Furthermore, of all the technically efficient points along the frontier QQ, point 1 is the most allocatively efficient way of producing quantity Q. Point 2 is as technically efficient as 1, but is less allocatively efficient, since production at 2 requires a budget of B_2, higher than B_1. Graphically, the allocatively efficient point (point 1) corresponds to the tangency between the budget constraint and the production possibilities frontier. Thus, technical efficiency is a pre-requisite for allocative efficiency. In general, two types of circumstances, discussed above, can lead to allocative inefficiency: technical inefficiency and technically efficient production that uses a mix of inputs that is not cost minimizing.[4]

Figure 2: Technical and allocative efficiency



Finally, when taken together, technical efficiency and allocative efficiency determine the degree of *economic efficiency*. Thus, if a DMU uses its resources in a technically and allocatively efficient way, then it can be said to have achieved economic efficiency.

4. There is a third cause of economic inefficiency, referred to as *social economic inefficiency* that can arise when the input prices faced by facility managers (for example, personnel wages or pharmaceutical products) depart from *social* (or *shadow*) prices.

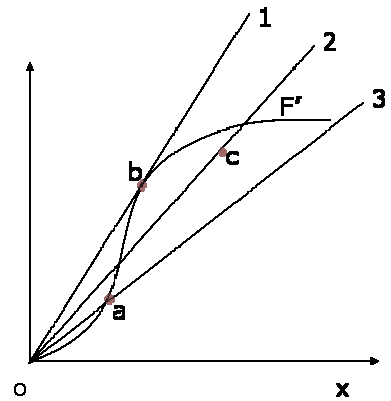
Alternatively, to the extent that either technical or allocative inefficiency is present, the DMU will be operating at less than total economic efficiency.

Productivity and efficiency

The productivity of a DMU is the ratio of the outputs(s) that it produces to the input(s) that it uses. Total factor productivity (TFP) is a productivity measure that involves all factors of production. Partial measures of productivity, such as labour productivity, on the other hand, involve only some factors. It is important to note that the terms productivity and efficiency are not synonyms; they describe two distinct, yet related, concepts. This implies that it is plausible for a DMU to experience high productivity while operating inefficiently (relative to other DMUs), and conversely, that an efficient DMU may experience low productivity (again, relative to other DMUs).

Figure 3 illustrates the difference between efficiency and productivity. Using y as output and x as input, OF' represents the production possibilities frontier, the convex portion of which reflects the presence of economies of scale. If a, b and c are DMUs such as hospitals, then only DMUs that lie on OF' can be regarded as technically efficient . Thus, while hospitals a and b are technically efficient, c is not. In addition, since the DMU with the highest ratio of y over x has the highest productivity, hospital b would be more productive than c, which is turn would be more productive than a.

Figure 3: The difference between productivity and efficiency



The next section provides a summary of the main empirical techniques used to estimate productivity and efficiency.

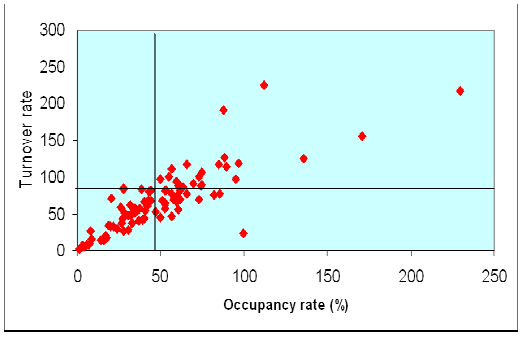
Alternative approaches to measuring productivity and efficiency

Adapting the classification of Barnum and Kutzin (1993), it is possible to categorise the main measurement approaches relevant to estimating productivity and efficiency. They identify three main approaches: input / output ratios and performance indicators; cost analyses; and statistical cost and production functions. To this list should be added frontier estimation methods.

Ratio measures

This is a simple way of measuring productivity. Input to output ratios approximate technical efficiency, and can help to quickly identify those facilities that are performing relatively poorly. Ratio measures have the advantage that they can usually be calculated using routinely collected data, e.g. the number of outpatient visits per doctor or nurse, or in the case of inpatient care, the average length of stay, bed occupancy or turnover rate. In addition, they are easy to estimate using data that tends to be readily available. In spite of these advantages however, ratio measures have limited utility largely because they generally focus on only one type of activity[5]. This is problematic given the multi-dimensional nature of health, and hospital services. Finally, inpatient and outpatient activities cannot be aggregated, and comparisons between hospitals offering different patterns of inpatient care are invalid.

Figure 4: Pabon-Lasso diagram



Source: Aparnaa Somanthan, Measuring health service productivity in developing countries: methods (Seminar 3)

Nevertheless, differences across both facilities, and between and within countries, can be compared through the use of the Pabon-Lasso diagram, which simultaneously presents data

5. Strictly speaking, input to output ratios are measures of partial coverage factor productivity, and as such they suffer from the fact that gains or losses in productivity may be imputed to the input in question when, in fact, they may result from changes to other inputs.

on length of stay, bed occupancy and turnover rates using sample means. An example is shown in Figure 4, using data from Sri Lanka. The diagram is divided into four quadrants. The southwest reflects capacity under-utilisation and low turnover, compared with the northwest quadrant, shows high utilisation and high turnover rates.

Unit cost analyses

One of the most commonly used techniques for measuring the costs of public health interventions is the accounting approach. Accounting cost studies provide unit cost estimates of, for example, admissions, bed-days, surgical procedures and outpatient visits, and as such approximate allocative efficiency. Thus they offer the potential for identifying low cost providers, which can be used as a benchmark against which to judge less efficient providers. Barnum and Kutzin (1993) divided accounting-based cost studies into two categories. The first uses detailed, bottom-up, step-down analyses of accounting to distribute shared costs across activities of individual facilities. The second uses a top-down approach, which makes less detailed estimates of high-level average costs based on aggregate expenditure records for multiple facilities.

Step-down costings typically attempt to assign costs to quite a low level of service provision (per department or specialty, perhaps even per procedure), by using various methods of allocating direct and overhead costs to a particular end-user department. Thus they tend to be detailed and resource-intensive. This inherently limits the number of units that can be examined in any given study. Clearly, the fewer the number of units to compare, the more difficult it becomes to make any judgments about relative efficiency. Aggregate data, by contrast, allows more scope for comparing relative performance in terms of average costs, but loses a significant degree of discrimination relative to step-down methods, since one can no longer differentiate resource use between different uses.

A problem common to both types of accounting studies is that they have an implicit underlying cost function represented by the sum of the products of the quantity of each input, multiplied by its respective price. Thus, although accounting studies generate a point estimate of total costs at an observed output, they do not provide information about what is likely to happen with changes in the price or quantity of an input. Inferences about economies of scale and scope therefore, cannot be made since average cost will only coincide with marginal cost under conditions of constant economies of scale.

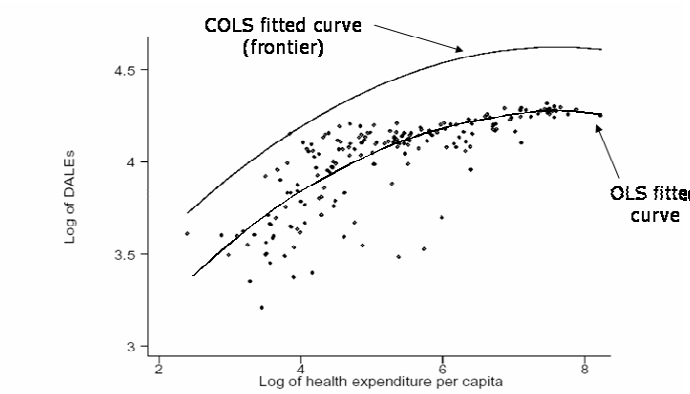
Statistical methods

Statistical methods use the estimated residuals from OLS estimates of production or cost functions to measure inefficiency. 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 measurement error. Applied to output-oriented data, points lying above the line of best fit represent units performing above average, and points lying below the line of best fit represent units performing below average. The opposite interpretation applies in the case of cost functions. Thus, while it is accepted that the estimated functions derived from such studies do not represent 'efficient' production[6], this approach is valuable in identifying the behaviour of marginal costs at different output levels, and in drawing conclusions regarding the existence and importance of returns to scale.

Multiple regression analysis is used to explore large numbers of independent variables, whose potential impact on cost can therefore be estimated. Although this allows adjustments for case-mix factors, this approach is not without its problems. Crucially, as the true functional form is not known in advance, there is always the inherent risk of mis-specification, which may yield misleading results. Attempts to use more flexible functional forms sacrifice capability to adjust for case-mix or other independent variables. Another limitation stems from the fact that all positive deviations from the predicted cost of output are interpreted as inefficiency, which may not be the case.

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 (Figure 5).

Figure 5: OLS and corrected OLS curves



Source: Bruce Hollingsworth and Andrew Street, An introduction to measuring efficiency and productivity in health and health care (Seminar 1)

6. The use of central tendency techniques inherently produces an analysis of average performance, i.e. not even best performance amongst inefficient producers.

In some sense, this line represents the production frontier, although it is only defined by the performance of the DMUs 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.

Statistical models, in contrast to cost analyses, provide a more realistic depiction of how total costs change in response to differences in service mix, inputs, input prices and scale of operations. It therefore allows for substitution between inputs as their relative prices and marginal productivity change. Statistical techniques are also more comprehensive than ratios because they accommodate multiple outputs and inputs.

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 productivity and relative efficiency.

Frontier Approaches

Frontier estimation methods involve the estimation of an efficiency frontier (or envelopment surface) from an observed sample of data, based upon best performance within the sample. The efficiency of other facilities in the sample is defined relative to these best performers. Specifically, measurement of the deviation of individual DMUs from this frontier enables the calculation of relative efficiency scores and the computation of potential efficiency gains if all units could achieve best performance levels.

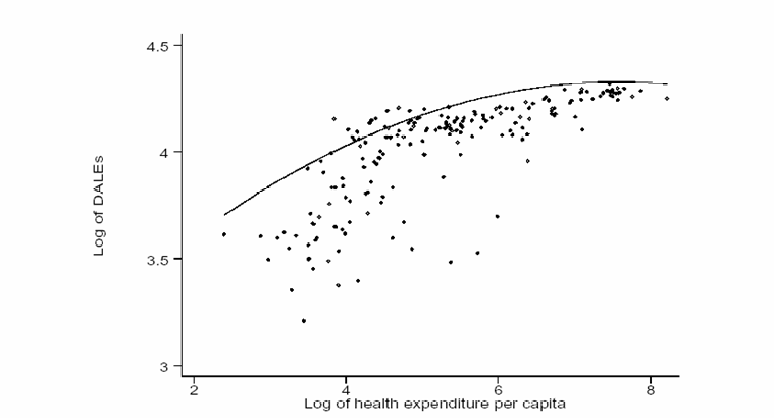
There are two major features that distinguish alternative empirical approaches for estimating the production frontier: whether they are parametric or not; and whether they are deterministic or stochastic. Parametric methods assume a specific functional form for the frontier, whereas non-parametric methods do not; and deterministic methods assume that the distance of a unit from its frontier is a result of inefficiency whereas stochastic methods assume that this is also partially due to random error.

SFA (Figure 6) 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, i.e. relative efficiency. SFA attributes any deviation from optimal performance

to either random or systematic sources of inefficiency by decomposing estimated residuals into the stochastic and systematic variations.

SFA's limitation stems from its reliance on untestable assumptions about the distribution of the error term, which is assumed to reflect inefficiency only. This increases the possibility of a specification error and also ignores random noise due to measurement errors and unobservable heterogeneity. SFA approaches also require large sample sizes and have more difficulty handling multiple outputs.

Figure 6: An example of SFA



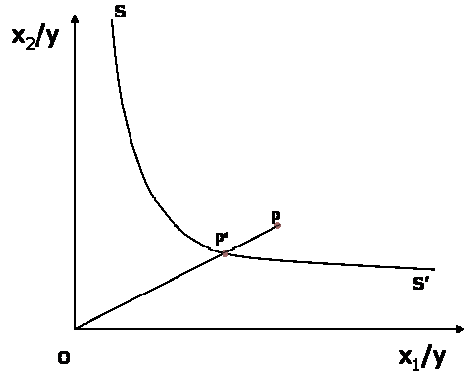
Source: Bruce Hollingsworth and Andrew Street, An introduction to measuring efficiency and productivity in health and health care (Seminar 1)

During recent decades, however, an alternative methodology to the stochastic frontier approach, DEA, has been developed and its application has grown rapidly over the years. It is a non-parametric approach. Thus is not subject to specification bias and is based on relative efficiency concepts proposed by Farrell (1957). In addition, DEA is a deterministic technique, which, as such, does not include explicitly a statistical error term reflecting measurement or sampling error. Farrell laid the foundation for new approaches to both efficiency and productivity studies at the micro level. Farrell's fundamental assumption was the possibility of inefficient operations, thereby pointing to a frontier production function concept as the benchmark, as opposed to a notion of average performance, which underlay most of the econometric literature on the production function up to the time of this seminal contribution. Charnes et al. (1978) extended and developed Farrell's approach for production units in the field of Operational Research.

In Figure 7, the DMU is producing a given level of output SS' using an input combination defined by point P. The same level of output could have been produced by radially

contracting the use of both inputs back to point P', which lies on the isoquant associated with the minimum level of inputs required to produce SS'. The input-oriented level of technical efficiency is defined by 0P' / 0P[7].

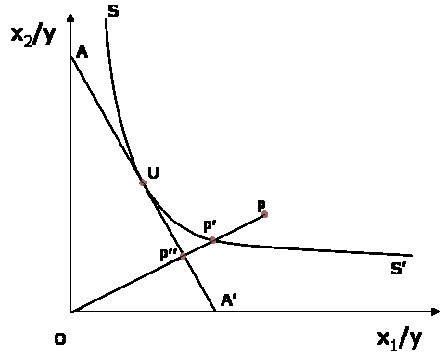
Figure 7: Input-oriented technical efficiency



Source: Bruce Hollingsworth and Andrew Street, An introduction to measuring efficiency and productivity in health and health care (Seminar 1)

In Figure 8 AA' represents an iso-cost line. Therefore, the least-cost combination of inputs that produces SS' is given by point U. To achieve the same level of cost (i.e. expenditure on inputs), the inputs would need to be further contracted to point P''. Allocative efficiency is therefore defined by 0P'' / 0P'.

Figure 8: Input-oriented technical, allocative and economic efficiency



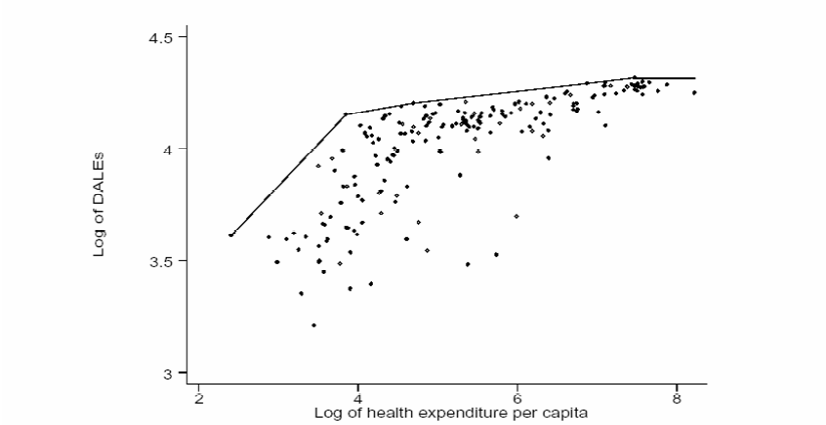
Source: Bruce Hollingsworth and Andrew Street, An introduction to measuring efficiency and productivity in health and health care (Seminar 1)

7. Efficiency can be considered in terms of the optimal combination of inputs to achieve a given level of output (an input-orientation), or the optimal output that could be produced given a set of inputs (an output-orientation). Thus the concept of efficiency can have naturally an output orientation or an input conserving orientation.

This means that (the economic efficiency of P) = (technical efficiency x allocative efficiency) = (OP' / OP) x (OP'' / OP') = (OP'' / OP). The point U is both technically and allocatively efficient, and therefore economically efficient as well.

DEA (Figure 9) employs linear programming to plot the extreme data points, which 'envelops' the data thereby creating the 'best practice frontier'. The efficiency of each provider is determined by its position relative to the frontier. 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. Another advantage is that it can handle multiple outputs as well as multiple inputs. Its disadvantage is that it estimates the efficiencies of the best performing units in that class and the entire residual (distance from frontier) is attributed to inefficiency.

Figure 9: An example of DEA



Source: Bruce Hollingsworth and Andrew Street, An introduction to measuring efficiency and productivity in health and health care (Seminar 1)

Stochastic estimations incorporate a measure for random error. This involves the estimation of a stochastic production frontier, where the output of a DMU is a function of a set of inputs, inefficiency and random error. An oft-quoted disadvantage of the technique, however, is that they impose both an explicit functional form, and distribution assumption on the data. In contrast, the linear programming technique of DEA does not impose any assumptions about functional form, and hence it is less prone to mis-specification. Furthermore, DEA is a non-parametric approach so does not take into account random error. Thus, it is not subsequently subject to the problems of assuming an underlying distribution about the error term. Furthermore, since DEA cannot account for such statistical noise, the efficiency estimates may be biased if the production process is largely characterised by stochastic elements.

Finally, SFA and DEA differ in the source of weights assigned to each output. Ideally, outputs would 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 each DMU is evaluated in the best possible light.

Issues in the measurement of efficiency in the production of health services

In this section, we review some of the methodological difficulties involved in efficiency analyses: adjusting for case mix; allowing for variation in technical quality; and knowledge of input prices. Finally some thoughts on choosing between the alternative approaches will be presented.

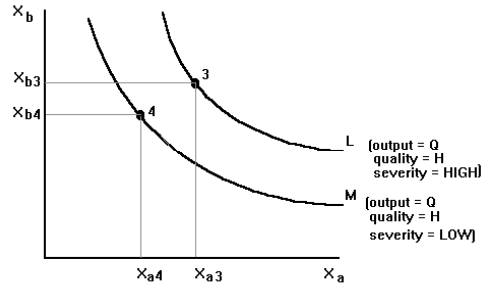
Efficiency and case mix

Case mix is an important, yet hard to define, concept through which researchers attempt to define hospital output. Available definitions involve some or all of the following terms: facilities (or services) available; intermediate and final services provided; complexity of the cases treated; and patient characteristics (for example, age and gender)[8]. Everything else being constant, one would expect efficient providers with different case mix to use different levels of inputs. For example, a facility with a greater proportion of complex cases should be expected to use more resources in producing health services to care for those cases, than an otherwise identical facility treating a set of patients with fewer severe cases.

Unless case mix is considered, comparative studies of technical and allocative efficiency among several providers are likely to be wrong. To illustrate this point, consider in Figure 10 the case of two providers, L and M, with L treating high severity patients, such as children with severe dehydration from dysentery, and M treating low severity patients, like children with mild dehydration from dysentery. Highly dehydrated children may need to remain hospitalised for several days, often receive intravenous feeding and rehydration, and require close attention by the facility staff. Children with mild dehydration on the other hand, can be sent home with instructions to the parents on oral rehydration salts and the appropriate treatment for dysentery.

8. Health related groups (HRGs) and diagnostic related groups (DRGs) are examples of systems developed to better reflect case mix.

Figure 10: Case mix and efficiency



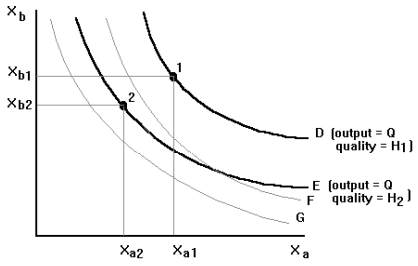
Suppose that provider L operated at point 3 to treat high severity cases while provider M operated at point 4 to treat the milder cases. If case severity was not considered, the uninformed researcher would wrongly conclude that provider M, the one with the lowest input use, is the more technically and allocatively efficient. If case mix were considered, however, the researcher would observe that the provider consuming the greatest amount of resources also happens to treat the most severe cases. Without further analysis, definitive statements about relative efficiency could not be made.

Efficiency and quality of care

Just as differences in case mix can obscure comparisons between technical and allocative efficiency among providers, so too can differences in the technical quality of care. Different levels of quality for example, often consume different levels of production inputs. Thus, failure to control for quality differences may ascribe higher efficiency to lower-quality producers and vice-versa.

To illustrate this point, consider the two providers in Figure 11, D and E, each capable of producing the same volume of output (e.g. Q ambulatory visits) according to their respective production possibilities frontiers. While both providers operate at the same output level, they produce different technical quality care: provider D is assumed to provide care of greater technical quality, H_1, while provider E is supposed to produce care of a lower technical quality, H_2.

Figure 11: Technical quality of care of efficiency



Suppose that provider D operates at point 1 and provider E operates at point 2. If an analyst attempting to compare technical and allocative efficiency between the two providers did not take into account their differences in technical quality, s/he would reach the conclusion that provider E is technically and allocatively more efficient than D. This would arise from the fact that provider E uses fewer production inputs than D (X_{a2} and X_{b2} versus X_{a1} and X_{b1}, respectively) and, as a consequence, provider E produces the level of output Q at a lower total cost than D. This conclusion however, would be wrong.

An appropriate comparison of efficiency is one which, at any given level of output, relates technical quality to input use. The analyst should therefore establish a relationship between H1 and (X_{a1}, X_{b1}) for provider D and compare it with the equivalent relationship between H_2 and (X_{a2}, X_{b2}) for the provider E. Given that the input levels X_a and X_b are in physical units however, it would be difficult to establish a quantitative relationship between health outcome levels (H) and the input levels.

Contrary to what is suggested by isoquants D and E in Figure 11, higher technical quality does not necessarily imply greater use of inputs. Although it is assumed that technical quality is higher along the isoquant D than along E, and also that resource use is greater for D, this does not necessarily have to be the case for all situations. For example, consider production of quantity Q according to F. Provider F's technical quality could be higher than E's, with F using smaller quantities of inputs when both providers operate at the far right of their possibilities (that is, production that is intensive in resource X_a). Alternatively, the technical quality of provider G could be greater than that of E at all points, yet with G consuming fewer inputs than E and thus being technically and allocatively more efficient.

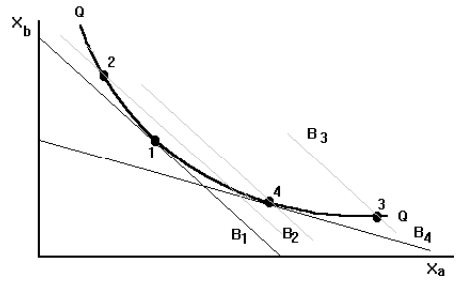
Allocative efficiency and input prices

Depending on a variety of circumstances such as the incentives, constraints, and information available to health facility managers, some providers may operate in a technically efficient, yet allocatively inefficient manner. For example, in the case of production input prices, allocative inefficiency arises when production occurs at a point that is not cost minimising. This can happen because facility managers either do not know their input prices or in spite of knowing the prices they fail to minimise their costs for a number of other reasons.

To distinguish between those two cases, consider the example of two providers operating at points 1 and 2 in Figure 12, each producing output level Q according to the same production possibilities frontier. Assume also that the two providers pay the same prices for their production inputs, X_a and X_b. Under those circumstances, provider 1 would be the most

allocatively efficient of the two because the production cost would be B_1, lower than B_2. If an analyst wanting to study the allocative efficiency for these providers knew that both face the same input prices, s/he would not need to measure those prices at all to rightly conclude that 1 is more economically efficient than 2.

Figure 12: Allocative efficiency and input prices



Suppose, instead, that providers 1 and 2 face different input prices. Unless the analyst knew exactly what those sets of prices were, s/he would be unable to make any statements about the providers' relative allocative efficiency. For example, although both providers could be cost minimisers, given the different prices that they face, they could also operate at different points along the production frontier. Alternatively, both could operate at points that are not cost minimizing. Thus, in order to ascertain relative allocative efficiency, both knowledge and the use of price information would be essential.

Conclusions

Theoretically, in order to measure the absolute technical and allocative efficiency of a production unit, one would need to know the underlying production and cost functions for that unit. This requirement poses significant problems for real-world application. First, the extreme heterogeneity and complexity of health care interventions (especially at the level of a large, multi-product production unit such as a hospital) effectively rules out the development of engineering-type production functions for all but the simplest interventions. If bottom-up engineering functions cannot be described, then some form of statistically derived estimation from observed data becomes necessary. In spite of this, one can only assume that a statistically estimated production or cost function reflects the underlying, 'true' function if one assumes that production units are always technically and allocatively efficient in their operation; there are good reasons to conclude that health care production units are unlikely to meet these conditions in reality. As a result, one must accept that the isoquant / isocost line of the efficient unit is unobservable, and any estimated production or cost function cannot be assumed to represent the production frontier or the underlying cost function (McGuire 1987).

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. In addition, it is necessary that the interpretation of the results obtained from all approaches should be tempered by an understanding of the chosen method's limitations. Furthermore, it is also important to define the theoretical framework underpinning the model and the reasons chosen for the model specification. Thus, in conclusion, it may be best to use a number of methods and then piece together the picture from the results. As Hollingsworth (2003) notes, "Given the limitations of frontier techniques at present it may be that they are best employed in tandem, when possible, and if different methods suggest similar directions for results then the validity of such findings is enhanced."

Review of data on productivity and efficiency

Several reviews have documented the growth in the literature, applying these techniques to various healthcare settings (Hollingsworth, 2003; Worthington 2004). For example, parametric and non-parametric methods have been used to examine the efficiency of a wide range of topics including: individual hospitals (Rosko 2001); primary care (Rollins et al. 2001); nursing homes (Hofler and Rungeling 1994); obstetric departments (Finkler and Wirtschafter 1998); dialysis (Ozgen and Ozcan 2002); stroke treatment (Ozcan et al. 1998); pharmacies (Capettini et al. 1985); and individual physicians (Chilingerian and Sherman, 1990).

While the literature has been predominantly concerned with the efficiency of US health institutions, applications from Austria (Hofmarcher et al. 2002), Australia (Hollingsworth et al. 2002), Belgium (Beguin 2001), Canada (Gruca and Nath 2001), Finland (Linna and Häkkinen 1998), Greece (Giokas 2001), Netherlands (Kooreman 1994), Norway (Erlandsen and Førsund 2002), Portugal (Dismuke and Sena 1999), Spain (Wagstaff and López 1996), Sweden (Gerdtham et al. 1999), Taiwan (Chang 1998), Turkey (Sahin and Ozcan 2000) and the UK (Jacobs 2001) have also been published. The methods have also been applied in a number of developing countries such as Bangladesh (Valdmanis et al. 2003), Botswana (Ramanathan, Chandra and Thupeng 2003) and South Africa (Zere et al. 2001). In addition, some papers have examined the efficiency of health care systems (Puig-Junoy 1998; Hollingsworth and Wildman 2003).

Apart from focussing soley on efficiency, these methods have also been used to study a range of issues such as: economies of scale and scope (Dacosta and Lapierre 2003); the impact of certain healthcare reforms (Gerdtham et al. 1999); ownership types (Rosenman, Siddharthan and Ahern 1997); competition (Cellini, Pignataro and Rizzo 2000); mergers (Harris, Ozgen and Ozcan 2000); hospital closures (Lynch and Ozcan 1994); technology use (Puig-Junoy 1997); diversification of hospital output (Prior and Sola 2000); and issues such as the change to the GP contract, and GP prescribing patterns (Bates, Baines and Whynes 1996).

Intra- and inter-country efficiency comparisons

Given that findings can differ for a number of reasons including differences in case mix and levels of technical quality, as discussed above, together with model specification issues, estimation techniques and data availability and quality (Hollingsworth 2003), results from different studies may not be strictly comparable. Results may therefore only be valid for the units under investigation, and hence are not necessarily generalisable. Based on this, Hollingsworth (2003) urges caution when interpreting his attempt at a meta-type analysis of the DEA results, the main findings of which are presented below in Table 1 and 2[9].

Table 1: Summary statistics for hospital efficiency scores

	No.	Mean	Median	Std deviation	Minimum
For profit	4	0.801	0.855	0.130	0.61
Non-for-profit	11	0.824	0.874	0.115	0.60
Public	6	0.948	0.945	0.033	0.895
Defence / VA	5	0.898	0.920	0.052	0.82
Non-teaching	2	0.742	0.743	0.046	0.71
Teaching	2	0.710	0.710	0.085	0.65
Acute / general	24	0.840	0.852	0.086	0.65
Non-specified	14	0.850	0.861	0.101	0,70
All hospitals	*68*	*0.844*	*0.870*	*0.099*	*0.60*
USA hospitals	*48*	*0.834*	*0.860*	*0.104*	*0.60*
EU hospitals	*17*	*0.892*	*0.897*	*0.073*	*0.751*
Non - USA / EU	*3*	*0.799*	*0.74*	*0.116*	*0.724*

Source: Hollingsworth (2003)

Defence / VA = Department of Defence hospitals and Veteran's Administration units

USA = United States of America

EU = European Union

9. Due to methodological incompatibility and small numbers, Hollingsworth did not pool the data from the stochastic frontier analyses. Rather summaries of the results were provided individually.

Table 1 summarises the results for each hospital type. The mean efficiency across the whole sample of hospitals is 0.84 and the median is 0.87. The data suggests there is substantial intra- and inter-country variation in the efficiency of hospitals. In addition, the data implies that there is the potential for efficiency gains across all hospital types, although the greatest gains would appear to be available to teaching hospitals and the lowest gains to public hospitals. It is also interesting to note that according to this data public hospitals are more efficient than private hospitals (mean efficiency of 0.948 vs. 0.801), and European hospitals are more efficient than their American counterparts (0.892 vs. 0.834).

However, it is important to note that the data presented by Hollingsworth (2003) was not collected for the purpose of cross-country comparisons, but rather, the studies were performed in isolation. Therefore, the current state of knowledge about cross-country differences in health service productivity and efficiency is limited. Thus, there is an urgent need to expand the research agenda to determine the reasons for the differences inferred by Hollingsworth (2003), and the relative impact of different strategies and policy levers on productivity and efficiency. In particular, the role of institutions and culture, as well as financial and organisational factors, in the incentive structure governing manager and provider behaviour, needs to be better understood if inter-country comparisons are to be interpreted correctly and if best practice is to be applied successfully across countries.

Table 2: Summary statistics for general health efficiency scores

	No.	Mean	Median	Std deviation	Minimum
Care programme	2	0.623	0.623	0.032	0.60
Health districts USA	9	0.742	0.800	0.144	0.50
Health districts EU	4	0.839	0.838	0.040	0.80
Nursing homes USA	18	0.746	0.806	0.175	0.38
Nursing homes EU	4	0.765	0.750	0.079	0.70
Primary care USA	4	0.648	0.635	0.249	0.427
Primary care EU	5	0.817	0.790	0.117	0.675

Source: Hollingsworth (2003)

US = United States of America

EU = European Union

Despite a large and growing body of literature on the measurement of health facility costs in developing countries (Barnum and Kutzin 1990; Adam et al. 2003), the literature on the measurement of efficiency is scant. Indeed, poor data availability in developing countries is likely to increase the cost, while limiting the sophistication and predictive power, of efficiency analyses that could be conducted. Although utilisation statistics can often be obtained at the

central or facility level, information on input prices and, more generally, costs, is seldom available in routinely kept records. This implies that studies of health facility efficiency in developing countries will generally have an important data collection component, a factor that will heavily increase the overall costs of research. Further, because government facilities are generally subsidised from the central level, data collection efforts often have to combine facility data with information obtained from the central level.

While data limitations are undoubtedly one explanation for the lack of research in this area, the limited volume of work may be explained largely by the fact that measuring efficiency is intrinsically much harder than measuring costs. There is an emphasis, albeit weak, on hospital efficiency research in developing countries, which coincides with that in the developed world. This emphasis on government hospital efficiency can partly be explained by the fact that: hospitals account for the largest share of health care costs; governments tend to keep information on utilisation and costs, however inaccurate, in a uniform way, whereas private providers generally do not; the search for health care financing and delivery reform has focused on gauging and improving the performance of the public sector. Nevertheless, the lack of comparative studies on efficiency between government and private providers is surprising, in light of the growing, yet empirically unsupported pressures on the part of experts and donors, to promote public divestment of curative care services in favour of a growing private participation.

Efficiency comparisons over time

In order to assess whether, and the extent to which, productivity and / or efficiency has varied over time, the Malmquist index can be used. The Malmquist index is the mean of two indices, measuring the change in efficiency from one period to the next, allowing a breakdown of efficiency changes over time[10]. Hollingsworth (2003) documents 22 Malmquist analyses from eight countries and an international analysis comparing 19 countries. The analyses of intra-country variation of productivity over time were applied to hospitals, primary care, pharmacies, ophthalmology and diagnostic technologies. All the studies documented productivity improvements apart from the study from South Africa (Zere et al. 2001), in which productivity declined by 12% among 86 hospitals between 1992-93 due to technology regress.

Rannan-Eliya presented data at the third OPI seminar which illustrated that, although sparse, there is enough evidence to suggest that there has been sustained health service

10. See Hollingsworth et al. (1999) for further details.

productivity growth in some developing countries but not in others – there was statistical evidence of productivity growth in six countries, Botswana, Mauritius, Uganda, Sri Lanka and Hong Kong; and a statistically significant decline in only two, Bahrain and Swaziland. These findings suggest that there is little evidence to support the fixed productivity assumption adopted in international policy proposals. They also suggest that more attention on strategies to improve productivity, relative to resource mobilisation, would be beneficial. Both are important although there is a particular issue about the extent to which additional resources result in an increase in the volume or quality of services delivered if the incentives for improved efficiency and productivity are weak.

The key question that arises is: why does productivity growth differ between countries and, if the Sri Lanka example is anything to go by, within countries? It may be significant that the six best performing countries were all at one time British Crown Colonies. This does not necessarily support the superiority of the British colonial project but it does point to the possible importance of institutional history and raises an intriguing topic for future research. However, although some explanations for differences in productivity growth can be advanced, much more needs to be known about how cultural, institutional, social, organisational and managerial factors that play off against each other before strategies to improve productivity can be designed on the basis of evidence, and before good practices from one country can be adopted successfully by others.

Apart from these factors, the interpretation of productivity analysis also needs to take into account the trade-offs that 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. In this regard, 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.

Quo vadis?

Increasing pressures on health sector resources have stimulated interest over the last decade in health services productivity and efficiency, and in ways to improve it. In OECD countries, with large publicly funded health care 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 health care 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.

This monograph has defined key concepts in this debate, provided a review of the alternative efficiency measurement tools and summarised what is known about cross-sectional comparisons between healthcare production units within areas or countries, or between countries, and in variations in productivity over time.

While there has been a recent expansion in the number of efficiency evaluations, there remains a dearth of literature from low- and middle-income countries. The measurement of health service productivity in developing countries poses specific challenges in addition to more general measurement difficulties. Nevertheless, this is disappointing given the developing world's greater scarcity of financial resources, which results in the inefficient use of scarce resources exacting a much higher penalty in terms of foregone health benefits.

More generally, further research is necessary in order to better understand the determinants of efficiency. 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. In particular, the relationship between efficiency and quality is an important, unresolved topic. In the first place, the cost implications of meeting minimum quality standards are unknown since the link between quality and outcomes is unclear. Second, although there is increasing interest in quality, the focus has been largely on clinical quality improvement – quality as perceived by the patient and relationships between technical quality and productivity over time have been neglected. Although measures such as mortality rates have traditionally been used, these may be affected by demand side distortions. For example, mortality rates in public hospitals may be higher than in private hospitals simply because private hospitals will not accept patients with complications. It is therefore necessary to exercise care when using mortality rates as a measure of the relative efficiency of private and public health facilities.

From a priority-setting perspective, it is pertinent to ask whether knowledge of technical inefficiencies matters, i.e. do inefficiencies distort priorities. Of course, in order to answer such a question, one needs data across both health programmes, and health sectors. A related question is whether it is more efficient to let the inefficiencies continue to exist at revealed levels, rather than intervene to correct them. In short, when is it cost-effective to

implement an efficiency improvement programme? As a starting point, an evidence-base on the costs and effects of strategies to improve efficiency needs to be collated. Of course it will be important to recognise the context-specific nature of many strategies, but consideration should be given to whether and how a matrix can be developed to summarise certain scenarios.

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.

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.

References

Adam T, Evans DB, Murray CJ. Econometric estimation of country-specific hospital costs. Cost-Effectiveness and Resource Allocation. 2003; 26;1(1):3

Barnum H, Kutzin J. Public Hospitals in Developing Countries: Resource Use, Costs and Financing. Population and Human Resources Department, The World Bank: Washington DC. 1990.

Bates JM, Baines DL, Whynes DK. Measuring the efficiency of prescribing by general practitioners. Journal of the Operational Research Society 1996; 47(12): 1443-1451.

Beguin C. Nonparametric frontier model as a tool for exploratory analysis of hospital stays. Methods of Information in Medicine 2001; 40: 241-247.

Capettini RAD, Morey RC. Reimbursement rate setting for Medicaid prescription drugs based on relative efficiencies. Journal of Accounting and Public Policy 1985; 4: 83-110.

Cellini R, Pignataro G, Rizzo I. Competition and Efficiency in Health Care: An Analysis of the Italian Case. International Tax and Public Finance 2000; 7 (4-5): 503-19.

Chang HH. Determinants of hospital efficiency: the case of central government-owned hospital sin Taiwan, Omega. The International Journal of Management Science 1998; 26(2): 307-317.

Chilingerian JA, Sherman HD. Managing physician efficiency and effectiveness in providing hospital services. Health Services Management Research 1990; 3(1): 3-15.

Dacosta CI, Lapierre SD. Benchmarking as a tool for the improvement of health services' supply determinants. Health Services Management Research 2003; 16(4): 211-223.

Dismuke C, Sena V. Has DRG payment influenced the technical efficiency and productivity of diagnostic technologies in Portuguese public hospitals? An empirical analysis using parametric and non-parametric methods. Health Care Management Science 1999; 2: 107-116.

Erlandsen E, Førsund FR. Efficiency in the provision of municipal nursing and home-care services: the Norwegian experience, in: Efficiency in the Public Sector, ed. Fox KJ. Kluwer: Boston, 2002

Farrell MJ. The measurement of productive efficiency. Journal of the Royal Statistical Society 1957; 120, series A Part III: 253-281.

Finkler MD, Wirtschafter DD. Cost-effectiveness and data envelopment analysis. Health Care Management Review 1993; 18(3): 81-88

Gerdtham UG, Rehnberg C, Tambour M. The impact of internal markets on health care efficiency: evidence from health care reforms in Sweden. Applied Economics 1999; 31: 935-945.

Giokas DI. Greek hospitals: how well their resources are used. The international Journal of Management Science 2001; 29: 73-83.
Gruca TS, Nath D. The technical efficiency of hospitals under a single payer system: the case of Ontario community hospitals. Health Care Management Science 2001; 4: 91-101.
Harris J, Ozgen H, Ozcan Y. Do mergers enhance the performance of hospital efficiency? Journal of the Operational Research Society 2000; 51: 801-811.
Hofler RA, Rungeling B. US nursing homes: are they cost-efficient? Economics Letters 1994; 44: 301-305.
Hofmarcher MM, Paterson I, Riedel M. Measuring hospital efficiency in Austria: a DEA approach. Health Care Management Science 2002; 5: 7-14.
Hollingsworth B, Dawson P, Maniadakis P. Efficiency measurement of health care: a review of non-parametric methods and applications. Health Care Management Science 1999; 2(3): 161-172.
Hollingsworth B, Harris A, Gospodarevskaya E. The efficiency of immunization of infants by local government. Applied Economics 2002; 34: 2341-2345.
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.
Hollingsworth, B. Non-parametric and parametric applications measuring efficiency in health care. Health Care Management Science 2003; 6(4): 203-218.
Jacobs R. Alternative methods to examine hospital efficiency: data envelopment analysis and stochastic frontier analysis. Health Care Management Science 2001; 4: 103-115.
Kooreman P. Data envelopment analysis and parametric frontier estimation: complementary tools. Journal of Health Economics 1994; 13(3): 345-346.
Linna M, Häkkinen U. A comparative application of econometric frontier and DEA methods for assessing cost efficiency of Finnish hospital, in: Health, the Medical Profession and Regulation, ed. Zweifel Z. Kluwer: Boston. 1998
Lynch JR, Ozcan YA. Hospital closure: an efficiency analysis. Hospital and Health Services Administration 1994; 39(2): 225-230.
Ozcan YA, Watts J, Harris JM, Wogen SE. Provider experience and technical efficiency in the treatment of stroke patients: DEA approach. Journal of the Operational Research Society 1998; 49: 573-582.
Ozgen H, Ozcan YA. A national study of efficiency for dialysis centers: an examination of market competition and facility characteristics for production of multiple dialysis outputs. Health Services Research 2002; 37(3): 711-732.
Prior D, Sola M. Technical efficiency and economies of diversification in health care. Health Care Management Science 2000; 3: 299-307

Puig-Junoy J. Measuring technical efficiency of output quality in intensive care units. International Journal of Health Care Quality Assurance 1997; 10(2/3): 117-120, 121-124.
Puig-Junoy J. Measuring health production performance in the OECD. Applied Economics Letters 1998; 5: 255-259.
Ramanathan TV, Chandra KS, Thupeng WM. A comparison of the technical efficiencies of health districts and hospitals in Botswana. Development Southern Africa 2003; 20(2): 307-320.
Rollins J, Lee K, Xu Y, Ozcan YA. Longitudinal study of health maintenance organisation efficiency. Health Services Management Research 2001; 14: 249-262.
Rosenman R, Siddharthan K, Ahern M. Output efficiency of health maintenance organisations in Florida. Health Economics 1997; 6: 295-302.
Rosko MD. Cost efficiency of US hospitals: a stochastic frontier approach. Health Economics 2001; 10: 539-551.
Sahin I, Ozcan YA. Public sector hospital efficiency for provincial markets in Turkey. Journal of Medical Systems 2000; 24(6): 307-320.
Valdmanis V, Walker D, Fox-Rushby JA. Are vaccination sites in Bangladesh scale efficient? International Journal of Technology Assessment in Health Care 2003; 19(4): 692-697.
Wagstaff A, López G. Hospital costs in Catalonia: a stochastic frontier analysis. Applied Economic Letters 1996; 3: 471-474.
Worthington AC. Frontier efficiency measurement in health care: a review of empirical techniques and selected applications. Medical Care Research and Review 2004; 61(2): 135-170
Zere E, McIntyre D, Addison T. Technical efficiency and productivity of public sector hospitals in three South African provinces. South African Journal of Economics 2001; 69(2): 336-358