European Research Studies,
Volume XII, Issue (3), 2009

Health Care Services Performance Measurement: Theory, Methods and Empirical Evidence

N. Maniadakis,
Department of Economics, University of Piraeus, Karaoli and Dimitriou 80 Street, 185 34 Piraeus

N. Kotsopoulos,
Government Affairs, GlaxoSmithKline Greece, 266 Kifisias Avenue, 152 32 Halandri

P. Prezerakos,
Vocational Training Centre, Municipality of Athens, 5 Favierou and Mayer Street, 104 38 Athens

J. Yfantopoulos,
Department of Political Studies and Public Administration, University of Athens
19 Omirou Street, Athens

Abstract
Despite the growing international literature in the field of efficiency and productivity measurement there are very limited Greek applications partly due to inadequate and incomplete datasets. The aim of this article is to illustrate the main methodologies for health care services efficiency and productivity measurement, to present their strengths and weaknesses and to discuss the existing evidence from applications in other countries. Notwithstanding the fact that the related methodologies have been recently developed these methods may help practitioners and health care decisions makers in improving health care management in Greece.

1. Introduction

In most western countries, the costs of health care have shown substantial increases during the last four decades and it is expected that this trend will continue in the future. In the search to explain this increasing trend of health care expenditure the list of factors that have been investigated includes: over-insurance, cost increasing technology, aging of population, supplier induced demand and the relative price effect. Recently, it has been argued that inefficiency and low growth productivity are two factors that have contributed to this of health care costs. Therefore, during the last twenty years, health services efficiency and productivity measurement and analysis has been the focal point of research and there is a rapidly growing literature in this field.

Health services performance measurement can be used for many purposes. Firstly, it can be used as a performance and success indicator and as a managerial control tool. In this context the measurement of efficiency may subsequently put pressures on the producers to improve performance. In addition it may be used to investigate the factors that drive better performance and thus to improve the management and organization of producers. Furthermore, the scarcity of the resources spent for the provision of health care, necessitates that these should be allocated to those producers that maximize the output produced with the given money spent for health care. In this context efficiency measurement may be used to guide resource allocation decisions. In addition, it can be used to evaluate alternative health care measures and policies and to test different hypotheses as regards to the performance of the health care industry.

Given the size of the health care industry, even small improvements in the performance of health care providers may generate substantial resource savings. Hence, measuring and analyzing the efficiency and productivity has become an issue of great importance. The approaches developed and applied to the performance measurement and analyses are rooted either on the economic and econometric literature or on the management science and operational research literature. Thus, they are based on different assumptions and techniques and because of the special characteristics of each approach, each has strengths and weaknesses. This article presents the main approaches proposed and used so far for health care services performance measurement. The evidence provided from the empirical literature is also reviewed.

2. Theory

2.1 Efficiency and its Measurement

Efficiency is a relative concept and it is measured in relation to optimum performance. An organization is *pure technically efficient* if output is produced with the least amount of factor inputs. *Scale efficiency* occurs when an organization operates at the optimal size of production, which is the point of long term equilibrium i.e. the point of constant returns to scale. Pure technical and scale efficiency combined define *technical efficiency*. An organization is said to be *allocatively efficient* if inputs are employed in the correct proportions in terms of

their prices, to minimize production cost for given output. The concept of *overall efficiency* combines allocative and technical efficiency as it occurs when an organization is simultaneously pure technically, allocatively and scale efficient.

To measure efficiency, we would ideally need to know what constitutes best performance in a technical sense but this is not feasible. In hospitals for example it is impossible to know what is the maximum output that may be feasibly produced or what is the minimum amount of inputs that the hospital should use given the technology it employees. Instead, in empirical work a group of hospitals is studied so as to find which are the ones that produce more output for given inputs or less inputs to produce certain output. Such best practice hospitals are then used to form a benchmark or in other words a frontier against which we measure the efficiency of the hospitals under evaluation.

To illustrate this process graphically consider Figure 1 which for simplicity depicts the single input-output case. Figure 1 depicts six hospitals namely, H1 to H6. The input quantities used by each hospital are measured in the horizontal axis and the output produced in the vertical axis. Thus, Hospital H6 uses input g and produces output a, but as shown in the figure hospital H1 also produces output a with only e quantity of input. Hence, at this level of production hospital H1 is efficient in the pure technical sense whereas, hospital H6 is not. The *input pure technical efficiency* measure of hospital H6 can be quantified by the ratio oe/og. Suppose that this ratio was 0.75 this would imply that hospital H6 can produce output a by using only 75% of the actual amount of inputs it currently uses. The same situation is depicted in the case of hospital H5 relative to H2 and H4 relative to H3. Thus, hospitals H6, H5 and H4 are technically inefficient because they can reduce input usage as compared to their peers (H1, H2, and H3). In contrast, hospitals H1, H2 and H3 are all technically efficient and they form a short term *production frontier* which exhibits *variable returns to scale* or VRS for short. Every hospital that lies on that frontier is also efficient whereas every hospital that lies in the interior of it is technically inefficient, where the amount of inefficiency is defined in terms of the distance from the frontier.

It should be noted though that there is also a difference between the three hospitals on the above mentioned frontier. That is, hospital H2 is related to greater product to input ratio (e.g. has greater productivity) relative to H1 and H3. This is due to the fact that this hospital operates at the most productive scale of production (MPSP), which is related to the long term production frontier, which exhibits constant returns to scale. Hence, even if hospitals H1 and H3 are pure technically efficient, they are not scale efficient. Hospital H1 operates in an area of increasing returns to scale and this implies that this hospital could increase further its efficiency by increasing its scale of production. Thus, the relative *scale efficiency* of hospital H1 is measured by the ratio od/oe. Had this hospital been scale efficient, it would have to operate at the production point that relates to output a and input d, which has the same output-input relation as the production point that relates to hospital H2. In the same spirit, hospital H3 is scale inefficient because it operates in an area of decreasing returns to scale. This hospital can improve performance by reducing its scale of production. Hence, its scale efficiency is measured by the ratio oh/oj.

Finally, note the case of hospital H5 which is pure technically inefficient but scale efficient because it operates at the right scale of production as it does hospital H2. Hospitals H6 and H4 are both scale and pure technically inefficient.

The concepts of efficiency were represented above from an input oriented perspective, in the sense that output was considered as given and the analysis studied how hospitals can reduce input usage whilst producing that output. Alternatively, one may take the input as given and examine how hospitals could increase output. The analysis so far has been confined only to input and output quantities and nothing was mentioned about input prices which affect production costs. Suppose that there are two inputs e.g. doctors and hospital capital and that they only produce one output, e.g. inpatient cases. This is depicted in Figure 2. To facilitate the presentation, output has been standardized so that the figure depicts input used per unit of output produced. Again hospitals H1 to H3 are efficient and they form the CRS frontier, which is also in this case called *isoquant.* The VRS frontier has been omitted in order to simplify the analysis. Technical efficiency in this case is measured by the amount by which it is possible to reduce radially (proportionally) input usage while still being able to produce a unit of output. Hospital H4 used the same input mix with H3 but it uses excessive amount of both inputs relative to its peer. The technical efficiency of H4 can be measured by the ratio OH3/OH4. Suppose now that input prices are known. Let for instance annual salaries be the price of the labor input and operating expense per annum be the price of capital input. The sum of the product s of inputs to their prices gives the operating cost of each hospital. This is not always to the minimum possible. Take for instance the case of hospital H4. The observed production cost of that hospital is represented by the *isocost line* that goes through point H4. This is in excess of the cost that relates to hospital H1 which uses input mix such that the marginal rate of substitution between the two inputs equals their relative price ratio. Hospital H1 is both technically but also cost and overall efficient. The production cost of hospital H4 can be radially contracted until it reaches the one that relates to H1. In other words hospital H4 can produce a unit of output at much lower cost. Its overall (cost) efficiency can be measured by the ratio OA/OH4, since at A cost is the same as H1. The line that goes through H1A is the *minimum isocost line.* The overall inefficiency of hospital H4 is due to two factors. Firstly, because of technical inefficiency, i.e. excessive input usage. This form of inefficiency is captured by the input oriented measure of technical efficiency that has already been discussed. Secondly, it is caused due to allocative inefficiency i.e. due to the fact that it employed a wrong input mix in light of their prices. In our example this form of inefficiency is measured by the ratio OA/OH3. In other words it is the residual inefficiency that explains cost inefficiency not accounted by technical inefficiency.

In summary, when only input-output quantity data are available it is possible to measure and decompose only technical efficiency. Price data can make it possible also to measure overall efficiency and then decompose it into a technical and an allocative part. The analysis illustrated above can be repeated from an output point of view. Nonetheless, because it is difficult to value output in health care it is more meaningless to analyze how to reduce production input factors and costs given the

outputs produced. This analysis has its theoretical roots to the works of Debrue (1951), Koopmans (1951) and Farrell (1957). However, it became more popular, when methods for the measurement of these concepts were provided by Charnes et al. (1978) and Aigner et al. (1977) and Jondrow et al. (1982), and subsequently developed (see Färe et al., 1985; Fried et al., 1993; Färe and Primmont, 1995) by others. Nonetheless, the first applications to health care institutions are seen in the American literature in the mid 80's. Since then a rapid growth in their applications and use for decision and policy making and planning was observed (Hollingsworth et al., 1999). The analysis outlined above was static and involved only data of one period. Thus, it gives only a snapshot of hospital performance. If panel data are available one may repeat it and get a better view on productive performance. However, in that case it is possible that apart from efficiency i.e. the relative distance of a hospital from its frontier, it is possible that the frontier itself may shift. This is the case when productivity change measurement is more appropriate.

2. 2 Productivity and its Measurement

Productivity is defined as the ratio of an index of output produced over an index of the input used to produce it. In the case of one input and output this is an easily obtainable measure, however, in the multiple input-output case, an economically sensible aggregation method has to be used. The over time change of this measure is called *productivity change*. For many years economists were attributing productivity change to technical (or technological) change. Technical change reflects the impacts from the introduction of new techniques, treatments, medical equipment etc. Thus, in this context the two concepts become synonymous. Nonetheless, after the recent developments of the literature on efficiency it was argued that productivity change can be caused from efficiency change rather than technical change. In other words, output per input used to produce it can increase not because of shifts in the frontier but because of shifts of the production unit (hospitals in our case) from it. Thus, the productivity measurement literature has recently incorporated the efficiency measurement literature. A large number of economists still measuring productivity ignoring efficiency but in such cases the effects of technical change are confounded with the effects of efficiency change and thus the analysis may provide inaccurate insights into productivity and its root sources (Grosskopf, 1993).

Figure 3 represents the case where, under constant returns to scale, a single input is used to produce a single in time periods, t and t+1, respectively. Let as assume that there are some hospitals in each period which define the production frontier and thus, hospital H1 that is under evaluation shifts from point $H1^t$ in period t to point $H1^{t+1}$ in period t+1. The frontier itself shifts upwards in period t+1 to technical progress. As noted earlier productivity is the ratio of output to input. Denote output in period t as y^t and input as x^t and similarly in period t+1 as y^{t+1} and x^{t+1}. Productivity change then is the change in average product that is: $PC = (y^{t+1}/x^{t+1})/(y^t/x^t)$. However, input here is the inefficient one and it has to be corrected for this inefficiency. Denote TE^t and $TE^{t+!}$ the factor by which input in each period

has to be corrected for inefficiency. Productivity change then becomes: PC = $(y^{t+1}/x^{t+1}\ TE^{t+1})/(\ y^t/x^tTE^t)$. Then after some algebraic manipulation we take that lnPC = $(lny^{t+1}\text{-}lny^t) - (lnx^{t+1}\text{-}lnx^t) - (lnTE^{t+1}\text{-}\ lnTE^t)$. In other words this is saying that change in productivity is attributed to technical change (change in output minus change in input) and technical efficiency change.

In this case the efficiency of hospital H1 in period t can be measured by the ratio od/of and in period t+1 by the ratio oe/og. Efficiency change then is captured by the ratio (oe/og)/(od/of). This ratio indicates whether the hospital overtime becomes more efficient i.e. whether it catches up its production frontier. A score less than 1 would here indicate that the hospital becomes more efficient over time and a score more than 1 would indicate the opposite. Technical change is captured by the ratio oc/od at the production mix of period t and by the ratio oe/oh at the production mix of period t+1. Both of them multiplicatively can define productivity change. A very popular way of pursuing this analysis is the Malmquist productivity index. This was first adopted in productivity measurement by Caves et al. (1982) and it was decomposed into technical efficiency change and technical change by Färe et al. (1989), Maniadakis and Thanassoulis (1996, 2000, 2004) extended it further so as to additionally capture allocative efficiency change. It should be noted again that it is easy to extend this analysis to include production costs rather than quantities only. It is also easy to generalize it to the output oriented case. For obvious reasons such extensions are not presented here but the interested reader can find them in Färe, Grosskopf and Lovell, (1994), Fried, Lovell and Schmidt, (1993) and Coelli, Batesse and Rao (1998).

3. Methods of Efficiency and Productivity Measurement

There are two main approaches to efficiency and productivity measurement. The *econometric approach* employs models that account for random noise and errors in the data and they are often called *stochastic***.** Also, because they require a prior assumption about the functional form of the technology, they are often termed *parametric*. However, it is rather confusing to use the term parametric only when referring to the econometric approach, because some other models are also parametric. In particular, the second approach to efficiency and productivity measurement is based on *mathematical programming* and includes *parametric* and *non-parametric* models. As implied by the name, non-parametric models do not require any assumption about the functional form of the technology, but they require more general assumptions such as convexity and non-emptiness. Mathematical programming models are also called *deterministic* because they do not account for random noise or data errors. This implies that to use them one needs to be confident about the quality of the data set. The non-parametric mathematical programming approach is the most popular in hospital efficiency and productivity studies.

3.1 The non-parametric mathematical programming approach

The non-parametric mathematical programming approach frequently goes by the name *Data Envelopment Analysis* or *DEA* for short, attributed to Charnes,

Cooper and Rhodes (1978) who introduced it. This name derives from the features of the method which is piece-wise linear and literally envelops the production input-output set. DEA can easily handle multi inputs-output technologies and it can be applied to small data sets, given that there is a reasonable proportion between the input and output variables and the number of observations. DEA is based on activity analysis which since its introduction the relevant literature has been rapidly expanding and now contains over two thousand theoretical articles and applications (see Emrouznejad and Thanassoulis (1997). There is a plethora of non-parametric programming models applied in efficiency and productivity measurement. In this article the basic ones will be represented.

Let being assumes that in any time period t, there are j =1,...,J hospitals which are using a vector of inputs $x^t \in \Re_+^n$, to produce a vector of output $y^t \in \Re_+^m$. The k^{th} producer consumes amounts x^t_{kn} of input n (n = 1,...,N) to produce amounts y^t_{km} of output m, (m = 1,...,M). Then technical efficiency or its reciprocal distance function (Färe, Grosskopf and Lovell, 1994), is measured in terms of how far is an input-output bundle from the piece-wise boundary and technical change is measured in terms of the over time shifts of that boundary. Specifically, for *every* hospital, the input oriented measure of technical efficiency discussed earlier can be computed as follows:

$$TE_i^t(y^t, x^t) = \min_{\lambda} \lambda, \text{ s.t.: } \sum_{j=1}^{J} z_j y_{jm}^t \geq y_{km}^t, \sum_{j=1}^{J} z_j x_{jn}^t \leq \lambda x_{kn}^t, \quad z_j \geq 0 \qquad (1)$$

where s.t. stands for subject to and variable z is an intensity variable used here to form the convex combinations of inputs and outputs. This model compares the distance of hospital k from the frontier formed by a group of its peers (and convex combinations of them) which produce output at least as large as the output of hospital k and use input less than or equal to the input of hospital k. To compute the input oriented measure of pure technical efficiency one has simply to add in above model the constraint: $\sum_{1}^{J} z_j = 1$, which makes the boundary of the technology to exhibit VRS. Then, the ratio of the two measures will give the scale efficiency measure. Assume also that in time period t input prices $w^t \in \Re_+^n$ are also available. This implies that the k^{th} hospital will face prices w_{kn}^t (n = 1,,N). For *every* hospital k, the minimum cost used to compute overall efficiency as follows:

$$C^t(y^t, w^t) = \min_{x,z} w_{kn}^t x \text{ s.t.: } \sum_{j=1}^{J} z_j y_{jm}^t \geq y_{km}^t, \sum_{j=1}^{J} z_j x_{jn}^t \leq x_n^t, z_j \geq 0, x_n \geq 0 \qquad (2)$$

This model is in the spirit of the one in (1) but it minimizes over inputs until it finds the minimum combination capable of securing the observed output. Note that this minimization may not necessarily be radial. For the computation of the input oriented overall efficiency measure one has simply to compute for every unit k, the

ratio: $C^t(y^t,w^t)/ \sum_1^n w_{kn}^t x_{kn}^t$, where the cost in numerator computed as in (2). Then, it is also possible to compute residually the input oriented allocative efficiency measure as the ratio of overall to technical efficiency. So far the focus was on efficiency measurement. To compute technical change or productivity change there are various alternatives. One option is it to use the so called *DEA Window Analysis*. This is an application of DEA to panel data. Assuming that there are t = 1,…,T years. Then, starting from the year t, "windows" of s years (s < T) are treated as a cross-sections and DEA is applied consecutively (see Charnes *et al.* 1995). However, this is an ad hoc method and not very much in line with the way technical change and productivity change measurement was defined in the previous chapters. An alternative option is to compute *measures of progress or regress.* These are described in Tulkens and Van de Eeckat (1995). There are many computational and conceptual difficulties in using such measures. Technical change and productivity change is measured in a much more elegant fashion with in the Malmquist index approach (Malmquist, 1953; Färe et al., 1989). To compute the index one needs to use models such as the one in (1).

3.2 The parametric mathematical programming approach

The parametric mathematical programming (PMP) approach requires the prior specification of a functional form for the technology of production. Then, the parameters of the model are calculated with the application of mathematical programming techniques, which were firstly used by Farrell (1957) and then extended by Aigner an Chu (1968); Forsund and Jansen (1977) and Forsund and Hjalmarsson (1979), among others. There is a vital difference between the DEA and PMP that frequently goes unnoticed. In DEA the boundary is estimated J times one for each producer in the data set and it may differ from one producer to the other. This led Färe *et al.* (1994) to note that DEA estimates individual frontiers. In contrast, PMP – as well as the econometric approach presented next – estimates one frontier that corresponds to the whole industry. Efficiency and productivity are then computed by substitution of the data set into the industry frontier. Assume as earlier that there are j =1,...,J producers which are using inputs $x^t \in \Re_+^n$, to produce output $y^t \in \Re_+^m$. There are various functional representations that may be used to represent the technology of production, such as the production function or the cost function. Assume that the productivity change measurement. They employed a Translog frontier production function, which was calculated with a linear programming model of the following form:

$$\min \Sigma[f(x_i , \beta,t) - y_i], \text{ s.t. } f(x_i , \beta,t) - y_i \geq 0 , \beta \geq 0 \quad (3)$$

where, x and y stand for inputs and outputs, β represents the parameters of the function and t the time.

The model can contain additional constrains, imposing the form of the returns to scale, concavity, monotonicity or any other property of the technology. One first needs to estimate the parameters, $\alpha,\beta,\gamma,\delta$, of the above technology using mathematical (linear or quadratic) programming. Then, efficiency measures or distances for every input-output bundle (i.e. producer) are computed by substitution into the estimated function.

3.3 The econometric approach

The econometric approach is similar to the parametric mathematical approach presented earlier in the sense that it requires prior specification of the form of the production technology. However, it also requires assumptions about the distribution of efficiency and errors. In addition, one needs to assume independence between the variables. A function estimated using ordinary least squares or a similar regression technique represents the average and not the frontier technology. Thus, techniques that estimate frontier functions are more appropriate in measuring efficiency and productivity. To estimate a function that will represent a frontier technology, in econometric frontier analysis the residuals in the regression model are bounded to be one-sided. The one-sided error term indicates how much a producer lies below the estimated production, revenue or profit frontier or above the estimated cost frontier and thus it is being used to determine its (in)efficiency or in general its distance from it. The shift of the frontier itself is used to measure technical change. For the discussion to follow let it be considered again that, in any time period t, hospital i (i = 1,,I) uses inputs $x_{in}^t \in \Re_+^n$, available at prices $w_{in}^t \in \Re_+^n$, to produce output $y_{im}^t \in \Re_+^m$. In the single output multiple input case, define the production function and assume that it has a parametric form. The production frontier function will then be as follows:

$$y_i^t = \alpha^t + \beta_n^t x_{ni}^t + \gamma k - e_i^t, e_i^t \le 0 \qquad (4)$$

α^t, β_n^t and γ are the parameters to be estimated and k is a dummy variable representing time, for the case when panel data are available. This model has two versions. In early work the entire error term, e_i^t, was bounded from above, i.e. $e_i^t \le 0$, so as the force $y^t \le F^t(\alpha^t, \beta_n^t x^t)$. Therefore, the entire error term represents technical (in) efficiency and this is why efficiency measures here are defined with reference to a deterministic frontier. This approach was first suggested by Aigner and Chu (1968) who proposed that the parameters of the model can be measured with linear or quadratic programming. However, Afriat (1972) and Schmidt (1976) proposed that this model can be made amenable to statistical analysis and showed how to estimate it. In later work by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1978) the error term was decomposed as follows: $e_i^t = \upsilon_i^t - u_i^t$.

The first component, υ_i^t, is a two-sided independently and identically distributed component, capturing statistical noise and random shocks. This

component accounts for the effects of events that are not under the control of the production unit, such as luck, weather, strikes, epidemics, accidents and the effects of measurement errors and omitted variable's bias. The second part, u_i^t, is a non-positive disturbance term, which forces $y^t \le F^t(\alpha^t, \beta_n^t x^t) + \upsilon_i^t$, and thus is used to measure technical (in)efficiency. Thus, in this model (in) efficiency is measured relative to a stochastic frontier. Estimation techniques include: corrected ordinary least squares (COLS), modified ordinary least squares (MOLS), and maximum likelihood (ML) (Schmidt 1986, and Fried, Lovell and Schmidt, 1993). In most of the cases the error component is assumed to follow a half-normal distribution, although other distributions (gamma, truncated) have been used. After the estimation the disturbance term can be decomposed into the two parts according to the approach proposed by Jondrow, Lovell, Materov and Schmidt (1982) and Waldman (1982).

The limitation of the production function is that it is applicable only in the case of single output multiple input technologies. In the case of multi input-output technologies one may use a cost function. The advantage of the cost function is that incorporates multiple outputs easily and that it makes it possible to estimate and decompose overall efficiency. The frontier cost function has the following form:

$$C_i^t = \alpha^t + \beta_m^t y_{im}^t + \gamma_n^t w_{in}^t + \delta k + e_i^t, e_i^t \ge 0 \qquad (5)$$

In this case the efficiency disturbance term is non-negative, so that it forces $C_i^t \ge C^t(\alpha^t, \beta_m^t y_m^t, \gamma_n^t w_n^t)$. Estimation techniques are the same as for the estimation of the production frontier function. It is notable that in a single-equation cross-section framework one is able to estimate overall efficiency but not to decompose it. The decomposition of overall efficiency requires the use of a system of equations. For the cases where panel data are available, Bauer (1990) shows how to do it.

4. Empirical Applications

During the last decade, the methodologies described in the previous sections have been extensively employed to measure and analyze the productive performance of health care services. Most of the applications were motivated by the desire to measure efficiency and productivity and to investigate their relation to observable characteristics of efficient organizations, especially ownership and profit status (profit vs. non-for-profit). It is worth noting that health industry is a particularly interesting area of efficiency and productivity measurement. Unlike a firm, a health care institution is not always expected to be efficient (Wagstaff, 1989). There is no obvious reason why a doctor should choose to be efficient. In the theory of the firm efficiency is a simple corollary of utility maximizing behavior. Firms try to maximize profits or minimize costs. However, as Evans (1971) suggests, hospitals do not adhere to maximizing/minimizing behavior in the traditional neoclassical

sense. There is no evidence that the public health care sector is profit-maximizing. Moreover, one of the criticisms of the frontier approach, that there might be a systematic cross sample variation in the technologies, is not valid in the health care sector. Usually, health care institution of similar type and size employ similar production technologies.

Looking into the literature it becomes clear the mathematical programming literature is much more extensive than the econometric one (Holingsworth, Dawson and Maniadakis, 1999). Most studies are focusing on the measurement health care services technical efficiency (Zukerman, Hadley and Iezzoni (1994), Wagstaff (1989), Parkin and Hollingsworth (1997), Valdmanis (1990, 1992), Register and Bruning (1987), Banker, Conrad and Strauss (1986), Morey, Fine and Loree (1990), Boussofiane, Dyson and Thanassoulis (1991), Sherman (1984), Maindiratta (1990), Ozkan, Luke and Haksever (1992), Ozkan and Luke (1993), Burgess and Wilson (1993, 1996), Vitaliano and Toren (1994), Kooreman (1994), Kleinsorge and Karney (1992), Nyman and Bricker (1989), Nunamaker(1983), Chilingerain (1995), Grosskopf and Valdmanis (1987), Huang and McLaughlin (1989), Kamis Gould (1991), Pina and Tores (1992), Rosko (1990), Sexton et al. (1989), and Thanassoulis, Boussofiane and Dyson (1995). There are only few studies which focus on allocative and overall efficiency such as those by Morey, Fine and Loree (1990), Byrnes and Valdmanis (1995), Maniadakis and Thanassoulis (2000). Finally, it is clear that productivity has only very recently attracted the research interest and thus, there are only few applications, such as those based on Malmquist indexes by Färe et al. (1989, 1992), Burgess and Wilson (1993c, 1995), Maniadakis and Thanassoulis (2000) and Maniadakis, Hollingsworth and Thanassoulis (1999).

The econometric studies are far lesser and are based usually on cost functions, like for instance in Wagstaff (1989), Dor (1994), and Zuckerman, Hadley and L. Iezzoni (1994). The most commonly used functional form is the translog. The models are estimated with COLS or ML and in all the studies the inefficiency error term is assumed to be half-normal distributed and it is decomposed according to the methodology proposed by Jondrow et al. (1982). A commonly used strategy to control for the multiproduct nature and heterogeneity of health care production is to stratify the organizations under investigation (hospitals, nursing homes etc.) into similar groups and assume that the technology and the output-mix is reasonably constant within each group. Hence, organizations are grouped according to their size, locality, teaching status, ownership, and other observable characteristics. Two other approaches that have been used to control for the heterogeneous nature of output is the service-mix approach (grouping according to services available or delivered) and the case-mix approach (Specialty mix, ICD groupings, DRG's).

Because of the special characteristics of health industry and the difficulties in measuring the final output of the health care provision, there is a lot of controversy about the choice of the appropriate input and output variables. The final production outcome of this industry, that is “health improvements”, is heterogeneous, multiple and it does not occur in district units. Thus, it is difficult to measure and at the same time take into account the quality of the health care services output. Subsequently, a significant proportion of variability exists in the chosen input and output sets

between different studies. The most commonly inputs used include: number of staff (doctors, nurses, technical, administrative and other staff or staff hours in different activities or staff days per client), costs or expenditure (total costs, salaries, food costs, total inpatient costs etc.), infrastructure and materials (beds, drugs, supplies, dressings, capital charges, net plant assets) and services (nursing, ancillary, administrative, general services). The most commonly used outputs include: number of patients, cases treated patient days and admissions. These output measures are usually desegregated according to the department occurred (inpatient, outpatient, surgical, maternity), the status of the person (physically disable, required limited care, personal care or residential care), age and sex.

In some studies, in a second stage, efficiency scores are regressed against various explanatory variables, in order to analyze the relation between efficiency and organizational characteristics. Variables that have been used include: institutional characteristics, ownership, size, occupancy rate, teaching status, age, affiliation, and organizational complexity, type of management, method of funding, staff indices, patient indices, length of stay, and many others. Because of the differences in the employed methodologies the results of the studies are controversial and can not be generalized, since most of the times they are focusing on organizations operating in specific environments. They are discussed in more detail in Hollingsworth, Dawson and Maniadakis (1999).

5. Discussion and Conclusion

During the last decades, the measurement of health services efficiency and productivity has been in the focus of the research interest and there is already an extensive literature which reflects this growing interest. However, because of the special characteristics of the health care industry, research in this field should be contacted cautiously and the results of different studies should be interpreted and used carefully. The inability to measure the final output of the health care industry and the low quality of the available data limit the application of the two methods described in the earlier sections. As Newhouse (1994) notes, these two techniques work better when the product is homogeneous and unidimentional (for example kilowatt per hour) and not multiple and heterogeneous like in the health care field. The same author states that it is almost certain that health industry studies suffer from omitted variables (input, output) bias. The techniques used to overcome these problems are not satisfactory and often have been criticized. To complicate matters, the estimated results are highly sensitive to changes in the basic assumptions or specifications of the used models.

Because of the above reasons, health services performance studies should use disaggregated observational data and they should concentrate on homogeneous and small segments of the health care system. In this case the number of outputs and the inputs are fewer, well defined and more accurate measurable. Also, the transformation from input to output i.e. the production technology could be better studied is smaller and more homogeneous production units. By studying less complex operational units, the analysis becomes simpler and hence, the used

throughputs may be better proxies of the real outcome and the calculated efficiency scores may better estimates of the real efficiency of the organization. The results of hospital (or other large institutions) efficiency studies are ambiguous and changeable, since there is no representative measure of real hospital outcome. Perhaps this is the reason why hospital efficiency studies provide contradicted results.

The accuracy of the estimated efficiency measures depends strongly on the use of an appropriate and well specified model, the inclusion of the relevant inputs and outputs, the use of the relevant data. The choice of the appropriate model is an important methodological issue in performance measurement. Researchers seem to exclusively apply the one approach or the other in all of the studies they conduct. Science management scientists argue that DEA is a superior method, while econometricians argue the opposite. However, the two methods are supplementary and not competitive. Both approaches have advantages and disadvantages and the choice of the most appropriate estimation method should depend entirely on the type of organizations under investigation and the quality of the available data. DEA is a non parametric method and does not assume a functional form for the frontier. Hence, it can accommodate wide ranging behavior in applications. This method operates sufficiently even for small samples; however, measurement errors can bias the results. Thus, DEA should be used for applications having relatively small measurement errors and not meeting the classical assumptions, even when the sample size is small. The econometric approach requires strong assumptions about the functional form and the distribution of the error term and needs large sample sizes, especially when many variables are used in conjunction with a translog function. Theoretically, when the assumptions of SFA are met, this approach has the advantage that it accounts for the effects of random shocks and statistical noise. However, deviations from the assumptions about the functional form and the distributions of the error term can result in estimation errors. Thus, the econometric approach should be used when there is evidence that the classical assumptions are met and when there is evidence of measurement errors or effects from the organizations environment. The measurement error and the organizational environment are the decisive factors which need to be considered prior to the choice of the appropriate research model.

A public sector institution, like a hospital, produces services for which it is difficult to quantify outputs and identify inputs. Additionally, outcome measures are highly susceptible to random fluctuation and the data suffers from measurement errors. In such cases, DEA should be used only when there is strong evidence that there are no measurement errors or effects from the organizations environment and the input-output set well defined and measurable. In other cases, it might be more appropriate to use the econometric technique. Due to the special features of the health care industry, these two methods must be tested and developed further, in order to be able to provide reliable and useful results. We believe that at the present they are more useful in investigating the association of performance with managerial and incremental characteristics of the production units and in hypothesis testing, rather than in providing final statements about individual organizational efficiency

and productivity. So far, the mathematical programming approach has been more extensively used in health services efficiency measurement. This method has to be improved so as to be able to isolate the effects of measurement errors and random fluctuations. Hence, the development of a stochastic DEA should be a primary research objective in this field. Surprisingly, the empirical application of the econometric technique in health care industry is very limited. Here, the research effort should be directed on the estimation of productive frontiers and productive efficiency. Another research objective should be the use of panel data. As discussed earlier, the observation of an organization for more periods is more likely to provide more consistent performance measures. Research should also be directed towards allocative efficiency and productivity measurement and analysis. On a large extend, progress in this field depends on the availability of high quality disaggregated information. Thus, in short term the improvement of health information systems must be a primary health policy objective. Due to the relative size of the health care sector, increases in efficiency and in productivity will result in substantial savings of resources, which can be redirected to uncovered areas and unsatisfied health care needs. Thus, health care services performance measurement and analysis should be a priority in the research agenda and research in this filed should be highly promoted.

References

1. Afriat, S.N., 1972, "Efficiency estimation of production functions", *International Economic Review, 13(3)* 568-598.
2. Ainger, D. J., Lovell, C.A.K., and P. Schmidt, 1977, "Formulation and estimation of stochastic production frontier models". *Journal of Econometrics 6*, 21-37.
3. Ainger, D. J. and S.F. Chu, 1968, "On estimating the industry production function'', *American Economic Review 13*, 58, 226-239.
4. Banker, R. D., Conrad, R. F. and R. P Strauss, 1986, "A comparative application of data envelopment analysis and translog methods: an illustrative study of hospital production", *Management Science 32(1)*, 30-44.
5. Bauer, P. W., 1990, "Decomposing TFP growth in the presence of cost efficiency, non constant returns to scale, and technological progress", *Journal of Productivity Analysis 1(4)*, 287-300.
6. Boussofiane, R. G., Dyson, R. G., and E. Thanassoulis, 1991, "Using Data Envelopment Analysis to assess the efficiency of perinatal care provision in England", *Warwich Business School Research Papers no.5*.
7. Byrnes, P., and V. Valdmanis, 1995, "DEA in hospital management: Analysing technical and allocative efficiency", in Charnes A, Cooper WW, Lewin AY, Seiford LM ed. "Data Envelopment Analysis: Theory, Method and Process.
8. Burgess, J., James, F., Wilson, P.W., 1993a, Technical Efficiency in Veterans Administration Hospitals: in *The Measurement of Productive Efficiency* (Fried HO, Knox-Lovell C.A, and Schmidt S.S., Eds). New York: Oxford University Press.

9. Burgess, J.F. Jr, and P.W. Wilson, 1993b, *Decomposing Hospital Productivity Changes 1985-88: A Nonparametric Malmquist Approach*: CORE - Universite Catholique de Louvain
10. Burgess, J.F. and P.W. Wilson, 1996, Hospital ownership and technical inefficiency. *Management Science*, vol 42 (No.1): pp 110 -123.
11. Burgess, J.F. and P.W. Wilson, 1995, Decomposing Hospital Productivity Changes, 1985-1988: A Nonparametric Malmquist Approach. *The Journal of Productivity Analysis vol 6*, 343-363.
12. Burgess, J.F. Jr, and P.W. Wilson, 1993c, "Decomposing Hospital Productivity Changes 1985-88: A Nonparametric Malmquist Approach", *CORE - Universite Catholique de Louvain*.
13. Caves, D.W., Christensen, L. R., and W.E Diewert, 1982, "The economic theory of index numbers and the measurement of input, outputandproductivity",*Econometrica50(6)*, 1393-1414.
14. Charnes, A., Cooper, W. W. and E. Rhodes, 1978, "Measuring the efficiency of decision making units", *European Journal of Operational Research 3(4)*, 392-444.
15. Charnes, Cooper, Lewin and Seiford (eds.), 1995, "Data envelopment analysis: theory, method and process. Management science series" (Quorum Books, New York).
16. Chilingerian, J. A., 1995, "Evaluating physician efficiency in hospitals - a multivariate- analysis of best practices", *European Journal of Operational Research 80*, 548-574.
17. Coelli, T., Prasada Rao D.S. and G. E. Battese, 1998, "An Introduction to Efficiency and Productivity Analysis", (Kluwer Academic Publishers, Massachusetts).
18. Debreu, G., 1951, "The coefficient of resource utilization", *Econometrica 19*, 273-292.
19. Emrouznejad, A. and E. Thanassoulis, 1997, "An Extensive Bibliography of Data Envelopment Analysis (DEA)" Volumes I, II and III. Working Papers 244, 245 and 258, Research Bureau, Business School, University of Warwick, Coventry.
20. Evans, R. G., 1971, "Behavioral cost functions for hospitals*" Canadian Journal of Economics, 4*, 198-215.
21. Dor, A., 1994, "Non-minimum cost functions and the stochastic frontier: on applications to health care providers", *Journal of Health Economics 13*, 329-334.
22. Färe, R., Grosskopf, S., and C.A.K Lovell, 1985, "The Measurement of Efficiency of Production" (Kluwer-Nijhoff Publishing).
23. Färe, R., Grosskopf, S., Lindgren, B. and P. Roos, 1992, "Productivity changes in Swedish Pharmacies 1980-89. a non-parametric Malmquist approach", *Journal of Productivity Analysis 3*, 85-101.
24. Färe, R. and D. Primont, 1995, "Mutli-output production and duality: theory and applications" (Kluwer Academic Publishers).

25. Färe, R., Grosskopf, S., Lindgren, B. and P. Roos, 1989, "Productivity developments in Swedish hospitals: A Malmquist output index approach" Discussion Paper No 89-3, Southern Illinois University, Illinois.
26. Färe, R., Grosskopf, S. and C.A.K. Lovell, 1994, "*Production Frontiers*" (Cambridge University Press, Cambridge).
27. Farrell, M. J., 1957, "The measurement of productive efficiency", *Journal of Royal Statistical Society 120*, 253-290.
28. Forsund, F. and L. Hjalmarsson,1979, "Generalised Farrell measures of efficiency: an application to milk processing in Swedish dairy plants" *Economic Journal 89*, 274-315.
29. Forsund, F. R. and E. S. Jansen, 1977, "On estimating average and best practice homothetic production functions via cost functions" *International Economic Review, 18(2)*, 463-476.
30. Fried, H., Lovell, C.A.K and J. Schmidt, 1993, "The measurement of productive efficiency", (University Press, Oxford).
31. Grosskopf, S. (1993) Efficiency and Productivity in H.O. Fried, C.A.K. Lovell and S.S. Schmidt eds. *The Measurement of Productive Efficiency: Techniques and Applications,* Oxford University Press, New York, 160-194.
32. Grosskopf, S. and V. Valdmanis , 1987, "Measuring hospital performance: A non-parametric approach", *Journal of Health Economics 6*, 89-107.
33. Hollingsworth, B., Dawson, P. and N. Maniadakis, 1999, "Efficiency Measurement of Health Care: A Review of Non-Parametric Methods and Applications", *Journal of Health Care Management Science 2(3)*, 161-172.
34. Huang, Y.L. and C.P. McLaughlin, 1989, "Relative efficiency of rural primary health care: and application of data envelopment analysis" *Health Services Research 24(2)*, 143-158.
35. Jondrow, J., Lovell, C.A.K., Materov, I., and P. Schmidt, 1982, "On the estimation of technical inefficiency in the stochastic frontier production function model", *Journal of Econometrics 19*, 233-238.
36. Kamis Gould E., 1991, "A case-study in frontier production analysis - assessing the efficiency and effectiveness of New Jersey partial care, mental health programs", *Evaluation and Program Planning 14 (4)*, 385-390.
37. Kleinsorge, I. and D. Karney, 1992, "Management of Nursing homes using Data Envelopment Analysis", *Socio-economic Planning Sciences 2*6, 57-71.
38. Koopmans, T.C., 1951, "Activity Analysis of production and Allocation, Cowles Commission for Research in Economics, Monograph No 13. Wiley, New York.
39. Kooreman P., 1994, "Nursing home care in the Netherlands: a non parametric efficiency analysis", *Journal of Health Economics 13(3)*, 301-316.
40. Maindiratta, A., 1990, "Largest Size-Efficient Scale and Size Efficiencies Of Decision-Making Units In Data Envelopment Analysis" *Journal Of Econometrics vol 46 (1-2)*, 57-72.
41. Malmquist, S., 1953, "Index numbers and indifferent surfaces", *Trabajos de Estadistica 4*, 209-242.

42. Maniadakis, N. and E. Thanassoulis, 1996, "A Malmquist index approach to productivity change measurement allowing for allocative inefficiency", Warwick Bussiness School Research Paper No 230.
43. Maniadakis, N., Hollingsworth, B. and E. Thanassoulis, 1999, "The Impact of the Internal Market on Hospital Efficiency, Productivity and Service Quality, Special Issue: Strategic Issues in Health Care", *Journal of Health Care Management Science 2(2)*, 75-85.
44. Maniadakis, N. and E. Thanassoulis, 2004, "A Cost Malmquist Productivity Index. Special Issue: DEA and its Uses in Different Countries", *European Journal of Operations Research 154(2)*, 396-409.
45. Maniadakis, N. and E. Thanassoulis, 2000, "Assessing Productivity Changes in UK Hospitals Reflecting Technologies and Input Prices", *Applied Economics 32*, 1575-1589.
46. Meeussan, W. and J. Van den Broeck, 1978, "Efficiency estimation from Cobb-Douglas production functions with composed error" *International Economic Review 18*, 435-444.
47. Morey, R., Fine, D. and S. Lorree, 1990, "Comparing allocative efficiencies of hospitals, Omega" *The International Journal of Management Science 18*, 71-83.
48. Newhouse, J.P., 1994, Frontier estimation: how useful a tool for health economics? *Journal of Health Economics*, *Vol.13*, No.3, 317-322.
49. Nunamaker, T.R., 1983, "Measuring routine nursing service efficiency: A comparison of cost per patient day and data envelopment analysis models", *Health Services Research 18*, 183-205.
50. Nyman, J.A. and D. L. Bricker, 1989, "Profit incentives and technical efficiency in the production of nursing home care", *The Review of Economics and Statistics 56*, 586-594.
51. Ozkan, Y. A., Luke, R. D. and C. Haksever 1992, "Ownwrship and organisational performance: a comparison of technical efficiency across hospital types", *Medical Care 30*, 781-794.
52. Ozkan, Y. A. and R. D. Luke, 1993, "A national study of the efficiency of hospitals in urban markets", *Health Services Research 27*, 719-739.
53. Pina, V. and L. Torres, 1992, "Evaluating the efficiency of nonprofit organisations: an application of data envelopment analysis to the public health services", *Financial Accountabiltity and Management 8(3)*, 213-225.
54. Parkin, D. and B. Hollingsworth, 1997, "Measuring production efficiency of acute hospitals in Scotland 1991-1994: validity issues in data envelopment analysis" *Applied Economics*: *29(11)*, 1425-1438.
55. Register, C. A. and E. R Bruning, 1987, "Profit incentives and technical efficiency in the production of health care", *Southern Economic Journal 53(4)*, 899-914.
56. Rosko, MD, 1990, "Measuring technical efficiency in health care organizations", *Journal of Medical Systems 14(5)*, 307-22.
57. Schmidt, P., 1986, "Frontier production functions", *Econometric Reviews 4*, 289-328.

58. Schmidt, P., 1976, "On the statistical estimation of parametric frontier production functions", *Review of Economics and Statistics 58(2)*, 238-239.
59. Sexton, T.R., Leiken, A.M., Nolan, A.H., Liss, A., Hogan, A., and R.H. Silkman, 1989, "Evaluating the managerial efficiency of veterans administration medical centers using data envelopment analysis", *Medical Care 27(12)*, 1175-1188.
60. Sherman, H.D., 1984, "Hospital efficiency measurement and evaluation: empirical test of a new technique", *Medical Care 22*, 922-938.
61. Thanassoulis, E., Boussofiane, A., and R.G. Dyson, 1995, "Exploring output quality targets in the provision of perinatal care in England using data envelopment analysis", *European Journal of Operational Research 80*, 588-607.
62. Tulkens, H., and P. Vanden Eeckaut, 1995, "Non-parametric efficiency progress and regress measures for panel data: methodological aspects" *European Journal of Operational Research, 80*, 3, 474-499.
63. Valdmanis, V., 1990, "Owneship and technical efficiency of hospitals", *Medical Care 28(6)*, 552-560.
64. Valdmanis, V., 1992, "Sensitivity analysis for DEA models - an empirical example using Public vs NFP hospital", *Journal of Public Economics 48*, 185-205.
65. Vitaliano, D.F. and M. Toren, 1994, "Cost and efficiency in nursing homes: a stochastic frontier approach", *Journal of Health Economics 13(3)*, 281-300.
66. Wagstaff, A., 1989, "Estimating efficiency in the hospital sector - a comparison of 3 statistical cost frontier models", *Applied economics 21(5)*, 659-672.
67. Waldman, D.M., 1982, A stationery point for the stochastic frontier likelihood. *Journal* of *Econometrics 18*, 275-279.
68. Zuckerman, S., Hadley, J. and L. Iezzoni, 1994, "Measuring hospital efficiency with frontier cost functions", *Journal of Health Economics 13 (3)*, 255-280.

FIGURES

Figure 1: Input Oriented Efficiency Measurement and its Components

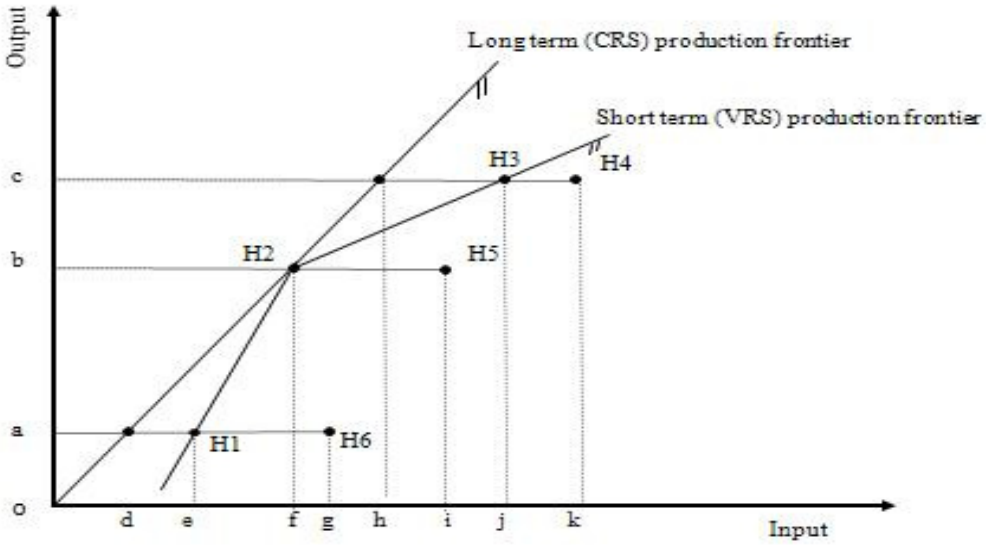

Figure 2: Input Overall (Cost) Efficiency and its Components

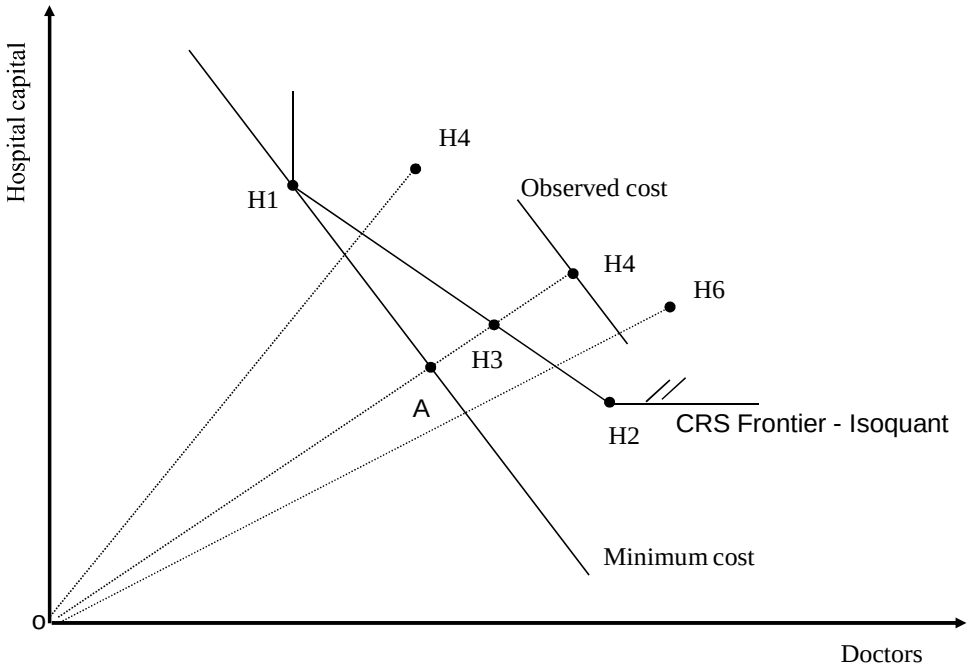

Figure 3: Input Oriented Productivity Measurement and its Components