HOSPITAL EFFICIENCY: CONCEPT, MEASUREMENT TECHNIQUES AND REVIEW OF HOSPITAL EFFICIENCY STUDIES

Hossein Moshiri, Syed Mohammed Aljunid, Rahmah Mohd Amin

1United Nation University- International Institute for Global Health, Kuala Lumpur, Malaysia
2Department of Community Health, Faculty of Medicine, University Kabangsaan Malaysia, Kuala Lumpur, Malaysia

ABSTRACT

In a time of rising demands on hospital reimbursement levels, focus on efficient operations is becoming more imperative. In health care systems, the measurement of efficiency is usually the first step in auditing individual performance of production units, e.g. hospitals, health centers, etc. It constitutes the rational framework for the distribution of human and other resources between and within health care facilities. The term efficiency is broadly used in economics and refers to the best utilization of resources in production. Typical example of efficiency is technical efficiency, referring to the effective use of resources in producing outputs. In the Farrell framework, a hospital is judged to be technically efficient if it is operating on the best practice production frontier in its hospital industry. In general, there are two main frontier methods in measuring efficiency. The first is Data Envelopment Analysis (DEA), a linear programming method which enables the measurement of efficiency consistent with the theoretically based concept of production efficiency. DEA typically examines the relationship between inputs to a production process and the outputs of that process. The second technique for assessing efficiency that is employed is Stochastic Frontier Analysis (SFA). This is an econometric technique to estimate a conventional function; with the difference being that efficiency is measured using the residuals from the estimated equation. The error term is therefore divided into a stochastic error term and a systematic inefficiency term.

Key words: Efficiency, Technical efficiency, Hospital Efficiency, DEA, SFA, Frontier method

INTRODUCTION

It is often argued that health care institutions are not expected to be efficient, as they do not adhere to neo-classical firm optimization behavior. However, given the vast amount of resources that goes towards funding such institutions, there is a great and growing interest in examining efficiency in hospitals with the driving force for such concern being value for money. Recently the demand for better quality health care services, accordingly the medical costs have been increased tremendously, which build a sharp contrast with very limited government resource and fund could be allocated to cop with this challenge. Increasing healthcare costs has been one of the most hotly debated policy issues in developed and developing countries in recent years. In many countries, public pressure and executive interest for cost control have led to various studies of the organizational causes of excess resource utilization; leading governments to seek new approaches to confront these critical issues.

Efficiency measurement represents a first step towards the evaluation of a coordinated health care system, and constitutes one of the basic means of audit for the rational distribution of human and economic resources. Over the past two decades, efficiency measurement has been one of the most intensely explored areas of health services research.
The aim of this paper is to provide a detailed review of the concept of efficiency, techniques of efficiency measurement, its application in hospital industry, and review some related studies.

The theory of microeconomic efficiency measurement
The recent history of microeconomic efficiency began in 1950 with Koopmans, who was the first formally defined technical efficiency. Debru (1951) first measured efficiency whereas Farrell (1957) who defined a simple measure of firm efficiency that could account for multiple inputs within the context of technical, allocative and productive efficiency.

There are some different components of economic efficiency; Pareto efficiency, Kaldor-Hicks efficiency, X efficiency. Pareto efficiency and Kaldor-Hicks efficiency are more philosophical concepts. The term ‘Pareto efficiency’ is named from Vilfredo Pareto, an Italian statistician and economist who used this term in his research of income distribution and economic efficiency. Given an alternative allocation for individuals, an allocation shift from one individual to another can make the former better without worsening the later. This is often called a Pareto optimization or Pareto improvement. The Kaldor-Hicks efficiency, named after Nicholas Kaldor and John Hicks, is another concept of economic efficiency that starts as an explanation of the limitation of unrealistic Pareto efficiency. Kaldor and Hicks’s concept of efficiency is more applicable to normal environment with less restricted criteria. X-efficiency, in contrast, is a more practical and measurable concepts. For example, Leibenstein’s X-efficiency means that if a company produces the maximum output, given available input resources such as workers, and machinery and technology, it is called X-efficiency.

In other classification, economists have developed three main measures of efficiency. First, technical efficiency; refers to the use of productive resources in the most technologically efficient manner. Put Technical efficiency implies the maximum possible output from a given set of inputs. Koopmans provided a formal definition of technical efficiency: “A producer is technically efficient if any output requires a reduction in at least one other output or an increase in at least one input, and if a reduction in any inputs requires an increase in at least one other input or a reduction in at least one output”. Within the context of healthcare services, technical efficiency may refers to the physical relationship between the resources allocated (capital, labor and equipment) and certain health outcomes. These health outcomes may either be defined in terms of intermediate outputs (number of patients treated, patient-days, waiting time, etc.) or final health outcomes (lower mortality rates, longer life expectancy, etc.). Second, allocative efficiency reflects the ability of an organization to use inputs in optimal proportions, given their respective prices and the production technology. In other words, allocative efficiency is concerned with choosing between the different technically efficient combinations of inputs used to produce the maximum possible outputs. Finally and when taken together, allocative efficiency and technical efficiency determine the degree of productive efficiency (also identified as total economic efficiency). Thus, if an organization utilizes its resources completely allocatively and technically efficiently, then it can be considered to have achieved total economic efficiency. Alternatively, to the extent that either allocative or technical inefficiency is present, then the organization will be operating at less than total economic efficiency.

Farrell (1957) described technical efficiency as the ratio of the firm’s observed output and the
maximum obtainable output on the frontier given observed factor utilization. The following figure illustrates Farrell arguments:

![Figure 1. Farrell efficiency figure](image)

We assume that there are two inputs, $x_1$ and $x_2$ used to produce a single output $y$. The production frontier is modeled as $y = f(x_1, x_2)$. Under the assumption of constant return to scale $1 = f(x_1/y, x_2/y)$. That is, as inputs increases, the relationship between the inputs and outputs does not change. The isoquant $SS'$ shows various combinations of two inputs that firm employs to produce a unit of output. The ratio $OQ/OP$ define the level of technical efficiency for a firm using inputs $(x_1^*, x_2^*)$ defined by point $P(x_1^*/y, x_2^*/y)$ to produce a unit of output, $y^*$. That is, point $Q$ represents an efficient firm using inputs $(x_1, x_2)$ in the same ratio as point $P$. Therefore, the ratio $OQ/OP$ measures the proportion of $(x_1, x_2)$ necessary to produce $y^*$. It follows that the ratio $OQ/OP$ measures the technical efficiency of the production unit of a firm operating at $P$. Therefore, $1 - OQ/OP$ measures the proportion by which $(x_1^*, x_2^*)$ could be reduced without reducing output. That is, it measures the possible reduction in the cost of producing $y^*$. Point $Q$ lies on the efficient isoquant. If the input price ratio that is represented by the slope of the isocost line $AA'$ is known, and then we can calculate the allocative efficiency (Farrell referred to as Price Efficiency). A ratio $OR/OQ$ indicates the production unit’s ability to use inputs in optimal proportions, given the respective prices at point $P$. Therefore $1 - OR/OP$ is the allocative inefficient point. The distance $RQ$ represents the reduction in production costs that would occur if production were to occur at the allocatively and technically efficient point $Q'$ rather than $Q$.

Therefore, if the production unit was perfectly efficient (both technically and allocatively), then the total economic or productive efficiency would be defined by the ratio $OR/OP$. The total inefficiency is therefore $1 - OR/OP$. We can interpret the distance $RP$ in terms of the cost reduction achieved by moving from the observed point $P$ to the cost minimizing point $Q'$.

Farrell (1957) defined these efficiency measures based on the assumption that the efficient production function is known. That is, he assumed that there are methods of comparing the observed performance of a production unit with some postulated standard. However, this is not usually the case. Therefore, efficient isoquant must be estimated using sample data.

The concept of hospital efficiency

In the Farrell (1957) framework, a hospital is judged to be technically efficient if it is operating on the best practice production frontier in its hospital industry. In the original Farrell framework, the entire observation on given sample is assumed to have access to same rechnology. Magnussen (1996) stated that measuring technical efficiency allows us to compare hospitals in terms of their real use of inputs and outputs rather than costs or profits. A hospital is said to be technically efficient if an increase in an output requires a decrease in at least one other output, or an increase in at least one input. Alternatively, a reduction in any input must require an increase in at least one other input or a decrease in at least one output. On the other hand allocative efficiency occurs when inputs or outputs are put to their
To measure hospital’s efficiency, the hospital’s output(s) must be identified. There are many potential measurements for a hospital’s outputs such as number of cases treated, number of procedures performed, number of patient days, bed turnover, and bed occupancy, among others. Which output or combination of outputs to use depends on the objectives of the hospital and on the level of measurement activities (e.g. departmental and institutional level).

**Approaches to study hospital efficiency**

Recent academic research on measuring efficiency in various areas has shifted to frontier efficiency. Frontier efficiency measures deviations in performance from that of best practice firms on the efficient frontier. In general there are two main approaches a nonparametric piecewise-linear convex isoquant constructed such that no observed point should lie to the left or below it (known as the mathematical programming approach to the construction of frontiers); or a parametric function, such as the Cobb-Douglas form, fitted to the data, again such that no observed point should lie to the left or below it (known as the econometric approach).

These methodologies estimate a best practice frontier with the efficiency of specific decision making unit measured relative to the frontier. The frontier efficiency of a firm measures how well that firm performs relative to the predicted performance of the best firms in the industry market conditions.

The econometric approach specifies a production function and normally recognises that deviation away from this given technology (as measured by the error term) is composed of two parts, one representing randomness (or statistical noise) and the other inefficiency. The usual assumption with the two-component error structure is that the inefficiencies follow an asymmetric half-normal distribution and the random errors are normally distributed. The random error term is generally thought to encompass all events outside the control of the firm, including both uncontrollable factors directly concerned with the ‘actual’ production function (such as differences in operating environments) and econometric errors (such as misspecification of the production function and measurement error). This type of reasoning has primarily led to the development of the ‘stochastic frontier approach’ which seeks to take these external factors into account when estimating the efficiency of real-world firm, and the earlier ‘deterministic frontier approach’ which assumes that all deviations from the estimated frontier represent inefficiency.

In contrast to the econometric approaches which attempt to determine the absolute economic efficiency of firm against, the mathematical programming approach seeks to evaluate the efficiency of a firm relative to other firms in the same industry. The most commonly employed version of this approach is a linear programming tool referred to as ‘data envelopment analysis’ (DEA).

Ferrier and Lovell (1990) illustrated that stochastic frontier analysis and data envelopment analysis may be used as crosscheck with each other12.

**Review of hospital efficiency studies**

In this section a review of selected efficiency studies regardless of purpose of study is provided. Valdmanis (1990) applied the DEA method to a group of hospitals and found that government-owned hospitals were more efficient13. This might be due to the fact that an imperfect adjustment is made for the quality of output and patient day rather than admission are generally used to measure output. The other surprising result is that for profit hospitals tend to be disproportionately represented among highly inefficient hospitals (Ozcan 1992) and are inefficient compared to not-for-profit hospitals when output is measure by discharging14.

Zuckerman, Hadley, and Iezzoni (1994) employed Stochastic Frontier Analysis for hospital cost functions15. Specifying a cost model that relied on input prices, output volumes, and output characteristics (i.e., less tangible output results such as the Joint Commission on the Accreditation of Health care Organizations score), the researchers determined that inefficiency accounts for (on average) 13.6 percent of total hospital costs, a result similar to Hofler and Folland.
Grosskopf and Valdmanis (1987) examined 22 public hospital and 60 private not-for-profit hospitals in California. They used DEA method and found that the two classes of hospitals to be facing distinct production frontiers with public hospitals being more efficient overall.

Zuckerman et al. (1994) used a cross-sectional stochastic frontier model to derive hospital-specific measure of inefficiency. The authors recognized that one of the goals of Medicare’s PPS in the US is to promote efficiency by rewarding hospitals that are able to keep their costs below PPS rates and penalizing those that are not. They also observed that a wide range of profitability among hospitals in 1990, which they attribute in part to the changes in the way that hospitals are paid. The existence of high profits for some hospitals and losses for others, lead the authors to question whether profitable institutions are efficient and those experiencing losses are not. If this is the case, it follows that inefficient hospitals should cut their costs and profitable hospitals should expand production (Zuckerman et al. 1994). Their stochastic frontier model measured the relative efficiency of hospitals so that they can better assess the relationship between profits and efficiency, thereby providing an answer to this question. According to their findings, the authors concluded that inefficiency accounts for 13.6 percent of total hospital costs and that the PPS which rewards efficiency and penalizes inefficiency, provides hospitals with appropriates incentives. This is because a reduction in inefficiency reduces costs. Their model showed that by removing the 13.6 percent estimated inefficiency this would have reduced hospital costs in the US in 1991 by approximately $ 31 billion. The findings also indicated some specific relationship with inefficiency.

One of the first applications of SFA to medical facilities (if not the first) was performed by Hofler and Folland (1991). Hofler and Folland suggested that SFA is important in assessing hospital costs and efficiencies, because other methods do not necessarily identify what minimum costs should be. The authors suggested that DEA is not entirely satisfactory because it ignores random fluctuations present in the data observations. In their research, Hofler and Folland assumed that structural cost differences based on ownership (for profit or not for profit categorization), teaching status, metropolitan or rural categorization, and Medicare volume ratio (high, low) were responsible for about 10.5% of hospital costs overall. The number of cost equations (12 equations based on the assumed differences in cost structure) and the irregularity of group size (as small as 35 and as large as 442) served to illustrate some of the problems associated with SFA: several equations could not be estimated, as the Maximum Likelihood Estimates did not converge.

Ozcan and Bannick (1994) used DEA to study trends in Department of Defense hospital efficiency from 1998-1999 using 124 military hospitals and data from the American Hospital Association Annual Survey. In a 1995 study, these authors also compared Department of Defense hospital efficiency with that of Veteran’s Administration hospital efficiency (n=284) using 1989 data. These studies were conducted at the strategic level under a different operational paradigm, prior to the large-scale adoption of managed care.

Charnes, et. al. (1985) conducted arguably the first Data Envelopment Analysis in military hospitals. Authors investigated the efficiency of 24 Army military hospitals during criteria that are still relevant for inclusion in their analysis. The authors selected traditional workload criteria for analysis of outputs including personnel trained, relative work product, and clinic visits. These outputs are considered traditional elements of production in health care and are relevant for inclusion along with other less traditional factors.

Ozcan and Luke (1993) used the DEA technique to conduct a national study of the efficiency of hospitals in urban markets. Four variables were analyzed in this study: hospital size, membership in multihospital system, ownership and payer mix. Ownership and percent Medicare were consistently related to hospital efficiency. The Medicare percent was related negatively to technical efficiency. Government hospitals were more efficient and for profit hospitals less efficient than other types of hospitals. Other variables like hospitals size, and membership in a multihospital system were related positively to efficiency.

Puig-Junoy (1998) used a cross-sectional DEA to study technical efficiency among ICUs in Spain.
using a two stage approach\textsuperscript{19}. In the first stage environmental factors, over which the ICU has no control, are ignored. In the second stage variation in operating efficiency was captured by a regression model. By focusing on the services provided by ICUs, the model alleviates the problem of measuring heterogeneous outputs, since all ICUs treat patients that are critically ill. Also analysis used patient-level data rather than aggregate data, and incorporates quality measures, such as mortality probability. Despite the emphasis on quality variables, the author acknowledged that the analysis does not attempt to measure whether patient receive an appropriate amount of care; rather it presents mortality probability data showing severity of illness at admission. Also, the outcomes for these patients are determined by survival status at discharge. The measurement of technical inefficiency requires that ICUs minimise inputs given the amount of outputs produced. The author acknowledged that measuring technical efficiency is adequate when comparing the performance of not-for-profit institution, such as those found in the hospital sector.

In 1998 Linna investigated the development of hospital cost efficiency and productivity in Finland by comparing both parametric and non-parametric panel models\textsuperscript{20}. The parametric panel methods has used stochastic frontier model with a time varying inefficiency component. The non-parametric panel methods use various DEA models to calculate efficiency scores and the Malmquist productivity index. Linna’s main objective in undertaking study was to determine if the use of panel data model would improve the estimates of individual efficiency scores compared to earlier cross-sectional analyses\textsuperscript{20}. The author found that results using panel data suggested that a reduction in inefficiency will reduce total hospital costs by between 1 and 1.2 billion Finnish Marks annually. These figures are slightly lower than those obtained using cross-sectional models, however the author noted that it is difficult to measure the significance of reliability improvement from cross-sectional data to using a panel. The results further indicated that the choice of modeling approach does not affect the results. SFA and DEA models were both able to reveal that productivity progress in 1988-1994 was due to both the exogenous rate of technical change and to the effect of time-varying efficiency. The author found that SFA and DEA methods produce different average efficiency score. Nevertheless, he concluded by saying that non-parametric and parametric methods used together with panel data provide a sufficiently clear understanding of the development of efficiency in hospital production to justify future studies of frontier models in health care.

Parkin and Hollingsworth (1997) used a constant return to scale to measure efficiency of 75 Scottish acute care hospitals\textsuperscript{21}. They use an input vector consisting of three capital and three labor variables and output vector consisting of four categories of inpatient discharges as well as emergency attendances and outpatient attendances. They found the rank correlation to range from 0.69 to 0.96

Another writer named as Craycraft (1999) recognized the necessity for non-profit organizations to measure efficiency due to the growing reliance, in particular in the US hospital sector, for government to base reimbursement on efficiency\textsuperscript{22}. The author noted that hospitals are reimbursed a fixed rate to compensate for efficient treatment. This author’s main concern was to show how important accurate efficiency measurement is in order to identify inefficiencies. Craycraft, reviewed various statistical techniques used in previous research to measure efficiency in hospitals and analyzed the strengths and weaknesses of each method\textsuperscript{22}. The techniques compared are: Ratio Analysis, Regression Analysis and Frontier Analysis (SFA and DEA). The author noted that measuring efficiency is difficult and inaccurate measures of efficiency may direct to poor decision. If efficiency is improperly measured, it may lead to a misallocation of resources among and within hospitals. If hospitals are considered inefficient when they are truly efficient, resources may be inappropriately allocated away from these hospitals. Craycraft's overviewed on the SFA technique set out its limitation when using cross-sectional data and promotes the use of panel data to overcome these limitations\textsuperscript{22}. Specifically the use of panel data overcomes the need to impose a functional form on the data. Clearly the SFA with panel data is superior in measuring relative efficiency because it overcomes the main objection to using a cross-sectional SFA, which is to impose a functional form on the data. Also, panel data
models require fewer assumptions because repeated observation on a number of decision making units, such as hospitals, can take the place of strong distributions assumptions\textsuperscript{23}.

Coppola (2003) conducted a DEA study of military hospital using 1998-2002 data. In his study he selected the following input variables: costs, number of beds, number of service offered\textsuperscript{24}. For output variables, he used surgical visit, ambulatory patient visit, emergency visits, and live birth. This study is focused on workload as the primary measure for efficiency, a point of view not fully congruent with the current operation of military hospitals.

Sherman (1984) wrote one of the founding articles on efficiency utilizing the DEA methodology on U.S. hospitals\textsuperscript{25}. He examined teaching hospitals and included nurses and interns trained as well as patient days as outputs. He compared and results of traditional ratio and regression analysis as well as DEA and found that DEA is a useful tool for the evaluation of resources among health care organizations and can lead toward improved hospital efficiency and reductions in health care costs. He suggested the DEA technique can overcome limitations of traditional and regression analysis and provide a more comprehensive measure of hospital efficiency

Mangnussen (1996) measured the production efficiency of 46 Norwegian hospitals using labor and capital inputs and specifying various output vectors\textsuperscript{11}. Notably, he examined treated patients and patient days as alternative units of measurement for inpatient activity and found the rank correlation between the models to be 0.67, implying substantial differences between the two measurement specification. As well, he examined the disaggregation of outputs based on patient complexity and the type of activity and found the rank correlation between the models to be 0.78 again revealing significant sensitivity to the model specification.

Street (2003) provided another application of SFA to the hospital sector using cross-sectional data for English public hospitals\textsuperscript{26}. More specifically the author compared the results obtained using corrected ordinary least square with results obtained using the SFA cost function. There are two alternative resulted obtained for the SFA model since the model is run under two assumptions of the distribution of the inefficiency term. One of the SFA models assumed a half-normal distribution, and the other an exponential distribution. Furthermore, the author produced confidence intervals relating to each hospital’s point estimate of relative efficiency. Findings from Street showed quite different levels of efficiency for each technique. The COLS model suggested that hospitals are on average 69 percent efficient, whilst the SFA model reported a mean efficiency of 90 percent. Although both models agreed on which hospital is the most efficient and which the least, the rate of efficiency varies, as does the ranking of hospitals in between these two extremes.

Another study is by Stanford’s (2004) examination of the performance by using DEA of 107 Alabama hospitals in the treatment of acute myocardial infarction patients because it too examined clinical efficiency and quality of care\textsuperscript{27}. Cross efficiencies were used to improve the efficiency discrimination between hospitals.

Bates (2006) used data envelopment analysis and multiple regression analysis to examine empirically the impact of various market-structure elements on the technical efficiency of the hospital services industry in various metropolitan areas of the United States\textsuperscript{28}. Market-structure elements include the degree of rivalry among hospitals, extent of HMO activity, and health insurer concentration. The DEA results showed the typical hospital services industry experienced 11 percent inefficiency in 1999. Moreover, multiple regression analysis indicated the level of technical efficiency varied directly across metropolitan hospital services industries in response to greater HMO activity and private health insurer concentration in the state. The analysis suggested the degree of rivalry among hospitals had no marginal effect on technical efficiency at the industry level. Evidence also implies that the presence of a state Certificate of Need law was not associated with a greater degree of inefficiency in the typical metropolitan hospital services industry
REFERENCE:


24. Coppola, M. N. 2003. *Correlates of military medical treatment facility (MTF) performance: measuring technical efficiency with the structural adaptation to regain fit (SARFIT) model and data envelopment analysis (DEA)*, A doctoral dissertation for a PhD in Health Service Organizations and Research, Virginia Commonwealth University, Richmond, VA.


