ASSESSING EFFICIENCY OF PRIMARY CARE SERVICES:
METHODOLOGICAL ISSUES

Saifulsyahira J.¹, Mohammad Farhan R.¹, Muhammad Hanafiah Juni²*

¹Doctor of Public Health Candidate (Health Service Management), Health Services Management Unit, Department of Community Health, Universiti Putra Malaysia
²Health Services Management Unit, Department of Community Health, Universiti Putra Malaysia

*Corresponding author: Associate Professor Dr Muhamad Hanafiah Juni
Health Services Management Unit, Department of Community Health, Universiti Putra Malaysia
Email: hanafiah_juni@upm.edu.my

ABSTRACT

The demand for high quality primary care services has been constantly increasing and it is likely to keep growing in the foreseeable future. Assessing efficiency of primary care services is crucial. The pursuit of efficiency in health systems should, therefore, be a central objective of decision makers and health managers. Measuring efficiency across institutions and across time is a critical element for improving the performance of health systems. The data for efficiency tool outlines key data categories and indicators necessary for assessing efficiency in the use of resource inputs, provides guidance on sources for these data and calculation of indicators. Applying this tool, managers will be better prepared to defend their budget requests providing evidence of internal efficiency while ensuring effective and efficient spending of monies that are allocated. There are a few methods that can be used for economic evaluation in assessing efficiency of primary care services. Economic efficiency of a service can be analyzed using non-parametric method (Data Envelopment Analysis) or parametric method - Stochastic Model (Frontier Approach). It can also be analyzed based on production function – Cobb-Douglas Model or Trans Log Model. Using General Additive Model (GAM) where additive modelling techniques are the impact of the predictive variables. These methodologies come with several issues, whether the methods are suitable or not to be used in assessing efficiency of primary care services. This article will disclose the issues and the application of the economic efficiency methods used in primary care services.

Keywords: Efficiency, Primary Care Services, Health, Methodologies, Issues.
1.0 INTRODUCTION

Primary care is the cornerstone of health care that is effective and efficient and meets the needs of patients, families, and communities. Strong primary care health systems are more likely to provide better population health, better distribution (more equity) in health throughout the populations, and greater economy in the use of resources. Kringos et al. (2010) reviewed 85 studies published between 2003 and 2008 for evidence in the areas of governance, economic conditions, workforce development, access, continuity of care, coordination of care, comprehensiveness of care, quality of care, efficiency of care and equity in health. The study found robust evidence that primary care contributes to overall health system performance and health (Starfield, 2012).

Benefits of primary care oriented health systems were consistent in showing greater effectiveness, efficiency, and equity. In the ensuing five years, nothing changed that conclusion, but there is now greater understanding of the mechanisms by which the benefits of primary care are achieved. Neither the wealth of a country nor the total number of health personnel are related to health levels. What counts is the existence of key features of health policy (Primary Health Care): universal financial coverage under government control or regulation, attempts to distribute resources equitably, comprehensiveness of services, and low or no co-payments for primary care services. All of these in combination, produce better primary care: greater first contact access and use, more person-focused care over time, greater range of services available and provided when needed and coordination of care. The evidence is no longer confined mainly to industrialized countries, as new studies show it to be the case in middle and lower income countries. Primary health care can now be measured and assessed; all innovations and enhancements in it must serve its essential features in order to be useful (Starfield, 2012).

Efficiency improvements in the health sector can yield considerable cost savings and even facilitate the expansion of services for the community. Minimizing waste, corruption and other forms of inefficiency, estimated between 20-40 percent of total health spending by the World Health Report 2010. These means that countries’ health systems can achieve more with the available resources. Why do some countries obtain relatively higher levels of service coverage and health outcomes than do others for the resources they invest? The answer lies in variation in efficiency. Inefficiencies exist everywhere in health systems with different degrees varying by country and setting. The most common sources of inefficiency include: inappropriate use of medicines; poorly executed procurement, including paying too much for medicines or technologies; misallocation and mismanagement of human and technical resources; underuse of capital equipment, particularly hospitals; excessive length of inpatient stays, or higher-than-needed admission rates; leakages, waste and corruption; medical errors; and an inappropriate mix of interventions (WHO, 2010).

The empirical applications of efficiency analysis were conducted in sectors like: accounting, advertising, auditing and law firms, airports, air transport, bank branches, bankruptcy prediction, community and health care, dentistry, education, electricity, environment, fishing, forestry, hospitals, hotels, macroeconomics, military, rail transport, sport, tax administration, water distribution etc. (Fried et al., 2008). The DEA (Data Envelopment Analysis) and SFA (Stochastic Frontier Approach) are the main methods commonly used to estimate efficiency of a DMU – Decision Making Units (commercial entities that produce tangible goods and
services that are sold in the market place, enterprise involved in delivering services or in the non-market sector, public bodies, national economic sector etc.) (Bezat, 2009).

2.0 METHODS FOR ASSESSING EFFICIENCY

2.1 Stochastic Model (Frontier approaches)

Stochastic frontier analysis (SFA) is a method of economic modeling. It has its starting point in the stochastic production frontier models simultaneously introduced by Aigner, Lovell and Schmidt (1977) and Meeusen and Van den Brink (1977). SFA employs multivariate statistical methods to explore output or cost variations between organizations and thereby produce efficiency scores for the entities under consideration (Lordan, 2007).

The production frontier model without random component can be written as:

\[ y_i = f(x_i; \beta) \cdot TE_i \]

where \( y_i \) is the observed scalar output of the producer \( i, i=1,...,I \), \( x_i \) is a vector of \( N \) inputs used by the producer \( i, f(x_i; \beta) \) is the production frontier, and \( \beta \) is a vector of technology parameters to be estimated.

\( TE_i \) denotes the technical efficiency defined as the ratio of observed output to maximum feasible output. \( TE_i = 1 \) shows that the \( i \)-th firm obtains the maximum feasible output, while \( TE_i < 1 \) provides a measure of the shortfall of the observed output from maximum feasible output.

A stochastic component that describes random shocks affecting the production process is added. These shocks are not directly attributable to the producer or the underlying technology. These shocks may come from weather changes, economic adversities or plain luck. We denote these effects with \( \exp \{v_i\} \). Each producer is facing a different shock, but we assume the shocks are random and they are described by a common distribution.

The stochastic production frontier will become:

\[ y_i = f(x_i; \beta) \cdot TE_i \cdot \exp \{v_i\} \]

Assume that \( TE_i \) is also a stochastic variable, with a specific distribution function, common to all producers. It can also be written it as an exponential \( TE_i = \exp \{-u_i\} \), where \( u_i \geq 0 \), since we required \( TE_i \leq 1 \). Thus, we obtain the following equation:

\[ y_i = f(x_i; \beta) \cdot \exp \{-u_i\} \cdot \exp \{v_i\} \]

Now, if we also assume that \( f(x_i; \beta) \) takes the log-linear Cobb-Douglas form, the model can be written as:

\[ \ln y_i = \beta_0 + \sum_n \beta_n \ln x_{ni} + v_i - u_i \]
Where \( v_i \) is the “noise” component, which we will almost always consider as a two-sided **normally distributed** variable, and \( u_i \) is the non-negative technical inefficiency component. Together they constitute a compound **error term**, with a specific distribution to be determined, hence the name of “composed error model” as is often referred.

Stochastic Frontier Analysis has examined also "cost" and "profit" efficiency (Kumbhakar & Lovel 2003). The "Cost frontier" approach attempts to measure how far from full-cost minimization (i.e. cost-efficiency) is the firm. Modelling-wise, the non-negative cost-inefficiency component is added rather than subtracted in the stochastic specification. "Profit frontier analysis" examines the case where producers are treated as profit-maximizes (both output and inputs should be decided by the firm) and not as cost-minimizers, (where level of output is considered as exogenously given). The specification here is similar with the "production frontier" one.

Stochastic Frontier Analysis has also been applied in micro data of consumer demand in an attempt to benchmark consumption and segment consumers. In a two-stage approach, a stochastic frontier model is estimated and subsequently deviations from the frontier are regressed on consumer characteristics (Baltas 2005).

A stochastic frontier production function is defined for panel data on firms, in which the non-negative technical inefficacy effects are assumed to be a function of firm-specific variables and time. The inefficiency effects are assumed to be independently distributed as truncations of normal distributions with constant variance, but with means which are a linear function of observable variables.

The second one widely uses stochastic procedure for parametric evaluating the frontier. The approach is stochastic – it considers additionally a random variable. The stochastic frontier approach treats deviations from production function as comprising both random error (white noise) and inefficiency. This enables a distinction between a random symmetrical component which accounts for measurement errors and stochastic effects (e.g. due to weather influences) and a symmetric deviation component which represents the inefficiency. The SFA as a parametric approach requires assuming a specific function form a priori, the frontier is estimated econometrically by some variant of last squares or maximum likelihood (Coelli et al. 2005). The SFA bases on econometric regression model, the frontier is smooth, appropriately and curved (Bezat, 2009).

Stochastic frontier models allow analysing technical inefficiency in the framework of production functions. Production units (firms, regions, countries, etc.) are assumed to produce according to a common technology, and reach the frontier when they produce the maximum possible output for a given set of inputs. Inefficiencies can be due to structural problems or market imperfections and other factors which cause countries to produce below their maximum attainable output.

Over time, production units can become less inefficient and catch up to the frontier. It is also possible that the frontier shifts, indicating technical progress. In addition, production units can move along the frontier by changing input quantities. Finally, there can be some combinations of these three effects. The stochastic frontier method allows decomposing growth into changes in input use, changes in technology and changes in efficiency, thus extending the widely used growth accounting method.
The standard definition of a production function is that it gives the maximum possible output for a given set of inputs; the production function therefore defines a boundary or a frontier. All the production units on the frontier will be fully efficient. Efficiency can be of two kinds: technical and allocative. Technical efficiency is defined either as producing the maximum level of output given inputs or as using the minimum level of inputs given output. Allocative efficiency occurs when the marginal rate of substitution between any of the inputs equals the corresponding input price ratio. If this equality is not satisfied, it means that the country is not using its inputs in the optimal proportions.

Polacheck & Yoon (1987) have introduced a three-component error structure, where one non-negative error term is added to, while the other is subtracted from, the zero-mean symmetric random disturbance. This modelling approach attempts to measure the impact of informational inefficiency on the prices of realized transactions, inefficiencies that in most cases characterize both parties in a transaction (hence the two inefficiency components, to disentangle the two effects).

Recently, various non-parametric and semi-parametric approaches were proposed in the literature, where no parametric assumption on the functional form of production relationship is made, see for example Parmeter and Kumbhakar (2014) and Park, Simar and Zelenyuk (2015) and references cited therein. 

The canonical formulation that serves as the foundation for other variations is their model,

\[ y = \beta^\prime x + v - u \]

Where \( y \) is the observed outcome (goal attainment), \( \beta^\prime x + v \) is the optimal, frontier goal (e.g., maximal production output or minimum cost) pursued by the individual, \( \beta^\prime x \) is the deterministic part of the frontier and \( v \sim N[0, \sigma_v^2] \) is the stochastic part. The two parts together constitute the „stochastic frontier.” the amount by which the observed individual fails to reach the optimum (the frontier) is \( u \), where

\[ u = |U| \text{ and } U \sim N[0, \sigma_u^2] \]

(Change to \( v + u \) for a stochastic cost frontier or any setting in which the optimum is a minimum). In this context, \( u \) is the inefficiency.” this is the normal-half normal model which forms the basic form of the stochastic frontier model.

2.2 Data-envelopment Analysis (DEA Model)

The first method is a non-parametric, deterministic procedure for evaluating the frontier. Non-parametric procedures determine a frontier which “envelops” the observations. DEA employs flexible, nonparametric methods to construct the best practice frontier and so allows the data to “speak for themselves” (Bezat, 2009). For the deterministic approaches the frontier is defined by the maximum distances between input and output. Random error and characterises deviations from the frontier are interpreted as inefficiency (Coelli et al. 2005). The DEA bases on a linear programming model which allows to build a piecewise linear frontier and assumes
a convex production sets. A less constrained alternative to DEA is non-stochastic method known as free-disposal hull (FDH) (Bezat, 2009).

Data Envelopment Analysis (DEA) is a decision making tool based on linear programming for measuring the relative efficiency of a set of comparable units. Besides the identification of relatively efficient and inefficient units, DEA identifies the sources and level of inefficiency for each of the inputs and outputs (Martic et al., 2009). DEA identifies a "frontier" which are characterized as an extreme point method that assumes that if a firm can produce a certain level of output utilizing specific input levels, another firm of equal scale should be capable of doing the same. The most efficient producers can form a 'composite producer', allowing the computation of an efficient solution for every level of input or output. Where there is no actual corresponding firm, 'virtual producers' are identified to make comparisons" (Berg 2010).

Data envelopment analysis (DEA) is a linear programming based technique for measuring the relative performance of organisational units where the presence of multiple inputs and outputs makes comparisons difficult.

According to DEA the efficiency of a multiple output, multiple input DMU $k$, with $k = 1, ..., n$, can be presented as follows:

$$\frac{\sum_j u_j y_{jk}}{\sum_i v_i x_{ik}} = \theta_k$$

Whereas $u$ measures the weight of each output $y_j$ ($j = 1, ..., s$), and $v$ indicates the weight of each input $x_i$ ($i=1, ..., m$) The efficient frontier of the group ($\theta = 1$) is constituted by the most efficient DMUs to which the efficiencies of the remaining DMUs are related.

To estimate the efficiency frontier, different important options have to be fixed. The orientation of the models reflects the appropriate direction of optimisation. This reflects which kind of quantities managers have better under control. In some branches, the organisations may be given a fixed quantity of resources and asked to produce as much output as possible. In this case the output oriented approach might be appropriate. By contrast, an input oriented approach should be used, if a fixed level of output has to be reached by using a minimal quantity of inputs. Because DMUs might differ according to their size and the quantities of inputs used and outputs produced respectively, assumptions concerning the returns to scale have to be formulated. The Charnes, Cooper, Rhodes (CCR) model incorporates constant returns to scale in production. The efficiency measure (1) and the usual side conditions can be adjusted accordingly. To obtain a linear programming problem, the Charnes Cooper transformation can be used. The output oriented linear programming envelopment for the DMU under evaluation $k$ is:
Whereas $\eta$ is defined based on the efficiency score of DMU $k$:

$$\eta_k = \frac{1}{\theta_k}$$

Vector $\varphi$ represents intensity variables which indicate the necessary combination of efficient entities (reference unit or peer) for every inefficient DMU in order to form a "virtual unit" or benchmark that is on the frontier.

Based on the CCR approach several other models were developed which build a profound basis for efficiency analysis with different returns to scale, different envelopment surfaces and different ways to project inefficient entities to the efficient frontier. Banker, Charnes and Cooper formulated the Banker, Charnes, Cooper (BCC) model which evaluates solutions for non-increasing returns to scale, non-decreasing returns to scale or variable returns to scale.

Whereas the CCR model only measures overall technical efficiency ($\theta_k^{CCR}$), the BCC model exclusively evaluates pure technical efficiency ($\theta_k^{CCR}$), because scale effects are taken into account. The comparison of CCR and BCC results enables to identify inefficiency which can be mitigated by increasing or decreasing the production volume resulting in a removal of scale inefficiencies. Thus, the ratio of the CCR and BCC efficiency measures will yield an estimate of the pure scale efficiency (SE) of DMU $k$, i.e.,

$$SE_k = \frac{\theta_k^{CCR}}{\theta_k^{BCC}},$$

indicating whether e.g. the different primary care facilities were operating on an efficient scale in producing their services. The optimal size of a DMU is reached when a marginal increase of all inputs (scale) leads to the same relative increase of outputs. The bigger the difference between the scale efficiency score of a DMU and full scale efficiency ($SE_k = 1$), the more unfavourable are the consequences of scale. It tells us how much output of a DMU can be expanded until it is as efficient as the reference unit (Marschall & Flessa, 2011).
2.3 Cobb-Douglas Model

In economics, the Cobb–Douglas production function is a particular functional form of the production function, widely used to represent the technological relationship between the amounts of two or more inputs, particularly physical capital and labour, and the amount of output that can be produced by those inputs. Sometimes the term has a more restricted meaning, requiring that the function display constant returns to scale (in which case $\beta = 1 - \alpha$ in the formula below). The Cobb-Douglas form was developed and tested against statistical evidence by Charles Cobb and Paul Douglas during 1927–1947.

In its most standard form for production of a single good with two factors, the function is,

$$ Y = AL^\beta K^\alpha $$

Where:
- $Y$ = total production (the real value of all goods produced in a year)
- $L$ = labour input (the total number of person-hours worked in a year)
- $K$ = capital input (the real value of all machinery, equipment, and buildings)
- $A$ = total factor productivity
- $\alpha$ and $\beta$ are the output elasticities of capital and labour, respectively. These values are constants determined by available technology.

Output elasticity measures the responsiveness of output to a change in levels of either labour or capital used in production, ceteris paribus. For example, if $\alpha = 0.45$, 1% increase in capital usage would lead to approximately a 0.45% increase in output.

Further, if $\alpha + \beta = 1$ the production function has constant returns to scale, meaning that doubling the usage of capital $K$ and labour $L$ will also double output $Y$.

If $\alpha + \beta < 1$ returns to scale are decreasing and if $\alpha + \beta > 1$ returns to scale are increasing.

Assuming perfect competition and $\alpha + \beta = 1$, $\alpha$ and $\beta$ can be shown to be capital's and labour’s shares of output.

Cobb and Douglas were influenced by statistical evidence that appeared to show that labour and capital shares of total output were constant over time in developed countries; they explained this by statistical fitting least-squares regression of their production function. There is now doubt over whether constancy over time exists.

The Cobb–Douglas function form can be estimated as a linear relationship using the following expression:

$$ \ln(Y) = a_0 + \sum_i a_i \ln(I_i) $$

Where:
- $Y$ = Output
- $I_i$ = Inputs
- $a_i$ = Model coefficients
The model can also be written as,

\[ Y = (I_1)^{a_1} \cdot (I_2)^{a_2} \cdots \]

As noted, the common Cobb–Douglas function used in macroeconomic modelling is,

\[ Y = K^\alpha L^\beta \]

Where \( K \) is capital and \( L \) is labour.

When the model exponents sum to one, the production function is first-order homogeneous, which implies constant returns to scale that is, if all inputs are scaled by a common factor greater than zero, output will be scaled by the same factor.

### 2.4 Trans Log Model

The Translog production function cannot be globally monotone, because there will be always a set of input quantities that result in negative marginal products. The Translog function would only be globally monotone, if all first-order coefficients are positive and all second-order coefficients are zero, which is equivalent to a Cobb-Douglas function. All Translog production functions fulfill the weak and the strong essentiality assumption, because as soon as a single input quantity approaches zero, the right-hand side of equation approaches minus infinity (if monotonicity is fulfilled), and thus, the output quantity \( y = \exp(\ln y) \) approaches zero. Hence, if a data set includes observations with a positive output quantity but at least one input quantity that is zero; strict essentiality cannot be fulfilled in the underlying true production technology so that the Translog production function is not a suitable functional form for analyzing this data set.

The input requirement sets derived from Translog production functions are always closed and non-empty. The Translog production function always returns finite, real, non-negative, and single values as long as all input quantities are strictly positive. All Translog production functions are continuous and twice-continuously differentiable.

\[
\ln q_i = \beta_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + 0.5\beta_{11}(\ln x_{1i})^2 + 0.5\beta_{22}(\ln x_{2i})^2 + \beta_{12}\ln x_{1i}\ln x_{2i} + \nu_i - u_i
\]

Production elasticity for \( i \)-th firm and \( j \)-th input is:

\[
E_{ji} = \beta_j + \beta_{j1}\ln x_{1i} + \beta_{j2}\ln x_{2i}
\]

Scale elasticity for \( i \)-th firm is:

\[
\varepsilon_i = E_{1i} + E_{2i}
\]

If we use transformed data where inputs are measured relative to their means, then Translog elasticities at means would simply be \( \beta_i \).
2.5 General Additive Model (GAM)

The GAM framework is based on an appealing and simple mental model:

\begin{enumerate}
  \item Relationships between the individual predictors and the dependent variable follow smooth patterns that can be linear or nonlinear.
  \item We can estimate these smooth relationships simultaneously and then predict \( g(\text{E}(Y))g(\text{E}(Y)) \) by simply adding them up.
\end{enumerate}

In linear regression, a linear least-squares fit is computed for a set of predictor or X variables, to predict a dependent Y variable. The well-known linear regression equation with m predictors, to predict a dependent variable Y, can be stated as:

\[ Y = b_0 + b_1 \times X_1 + \ldots + b_m \times X_m \]

Where \( Y \) stands for the (predicted values of the) dependent variable, \( X_1 \) through \( X_m \) represent the m values for the predictor variables, and \( b_0 \) and \( b_1 \) through \( b_m \) are the regression coefficients estimated by multiple regression. A generalization of the multiple regression model would be to maintain the additive nature of the model, but to replace the simple terms of the linear equation \( b_i \times X_i \) with \( f_i(X_i) \) where \( f_i \) is a non-parametric function of the predictor \( X_i \). In other words, instead of a single coefficient for each variable (additive term) in the model, in additive models an unspecified (non-parametric) function is estimated for each predictor, to achieve the best prediction of the dependent variable values.

The general linear model for a single dependent variable can be considered a special case of the generalized linear model: In the general linear model the dependent variable values are expected to follow the normal distribution and the link function is a simple identity function.

In the general linear model a response variable Y is linearly associated with values on the X variables while the relationship in the generalized linear model is assumed to be:

\[ Y = g(b_0 + b_1 \times X_1 + \ldots + b_m \times X_m) \]

Where \( g(\ldots) \) is a function. Formally, the inverse function of \( g(\ldots) \), say \( g_i(\ldots) \), is called the link function; so that:

\[ g_i(\mu Y) = b_0 + b_1 \times X_1 + \ldots + b_m \times X_m \]

where \( \mu Y \) stands for the expected value of \( Y \).

We can combine the notion of additive models with generalized linear models, to derive the notion of generalized additive models, as:

\[ g_i(\mu Y) = S(f_i(X_i)) \]

In other words, the purpose of generalized additive models is to maximize the quality of prediction of a dependent variable Y from various distributions, by estimating unspecific (non-parametric) functions of the predictor variables which are "connected" to the dependent variable via a link function.
Mathematically speaking, GAM is an additive modelling technique where the impact of the predictive variables is captured through smooth functions which depend on the underlying patterns in the data can be nonlinear:

We can write the GAM structure as:

\[ g(E(Y)) = \alpha + s_1(x_1) + \cdots + s_p(x_p), \]

where \( Y \) is the dependent variable (i.e., what we are trying to predict), \( E(Y) \) denotes the expected value, and \( g(Y) \) denotes the link function that links the expected value to the predictor variables \( x_1, \ldots, x_p \).

The terms \( s_1(x_1), \ldots, s_p(x_p) \) denote smooth, nonparametric functions. Note that, in the context of regression models, the terminology nonparametric means that the shape of predictor functions is fully determined by the data as opposed to parametric functions that are defined by a typically small set of parameters. This can allow for more flexible estimation of the underlying predictive patterns without knowing upfront what these patterns look like. For more details on how to create these smooth functions.

Note that GAMs can also contain parametric terms as well as two-dimensional smoothers. Moreover, like generalized linear models (GLM), GAM supports multiple link functions. For example, when \( Y \) is binary, we would use the logit link given by

\[ g(E(Y)) = \log \frac{P(Y=1)}{P(Y=0)} \]

### 3.0 ADVANTAGES AND DISADVANTAGES OF THE ECONOMIC MODELLING

<table>
<thead>
<tr>
<th>Methods</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Stochastic Model         | • random variations in catch can be accommodated, so that the measure is more consistent with the potential harvest under “normal” working conditions  
                          | • Provides an overall, objectively determined, numerical efficiency value and ranking of contracts that is not otherwise available. | • more complicated, requires stochastic multiple output distance functions, and raises problems for outputs that take zero values  
<pre><code>                      |                                                                          | • Large number of decision making units (DMUs).                           |
</code></pre>
<p>| Data-envelopment Analysis | • no need to explicitly specify a mathematical                            | • results are sensitive to the selection of inputs and                         |</p>
<table>
<thead>
<tr>
<th>Model</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEA Model</td>
<td>- Form proven to be useful in uncovering relationships that remain hidden for other methodologies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Capable of handling multiple inputs and outputs</td>
<td>- Cannot test for the best specification</td>
</tr>
<tr>
<td></td>
<td>- Capable of being used with any input-output measurement</td>
<td>- The number of efficient firms on the frontier tends to increase with the number of inputs and output variables (Berg 2010)</td>
</tr>
<tr>
<td></td>
<td>- The sources of inefficiency can be analysed and quantified for every evaluated unit</td>
<td>- It is sensible to outliers, error measurements and random influences in the data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Deems any deviation from the efficiency frontier to be the result of inefficiency.</td>
</tr>
<tr>
<td>Cobb-Douglas Model</td>
<td>- It suits to the nature of all industries.</td>
<td>- The function includes only two factors and neglects other inputs.</td>
</tr>
<tr>
<td></td>
<td>- It is convenient in international and inter-industry comparisons.</td>
<td>- The function assumes constant returns to scale.</td>
</tr>
<tr>
<td></td>
<td>- It is the most commonly used function in the field of econometrics.</td>
<td>- There is the problem of measurement of capital which takes only the quantity of capital available for production.</td>
</tr>
<tr>
<td></td>
<td>- It can be fitted to time series analysis and cross section analysis.</td>
<td>- The function assumes perfect competition in the factor market which is unrealistic.</td>
</tr>
<tr>
<td></td>
<td>- The function can be generalised in the case of ‘n’ factors of production.</td>
<td>- It does not fit to all industries.</td>
</tr>
<tr>
<td></td>
<td>- The unknown parameters ‘a’ and ‘p’ in the function can be easily computed.</td>
<td>- It is based on the substitutability of factors and neglects complementarity of factors.</td>
</tr>
<tr>
<td></td>
<td>- It becomes linear function in logarithm.</td>
<td>- The parameters cannot give proper and correct economic implication.</td>
</tr>
<tr>
<td></td>
<td>- It is more popular in empirical research.</td>
<td></td>
</tr>
<tr>
<td>Trans Log Model</td>
<td>- Flexible functional form</td>
<td>- More difficult to interpret</td>
</tr>
<tr>
<td></td>
<td>- Less restrictions on production elasticities and substitution elasticities</td>
<td>- Requires estimation of many parameters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Can suffer from curvature violations.</td>
</tr>
</tbody>
</table>
| General Additive Model | • Easy to interpret- Has the interpretability advantages of GLM are where the contribution of each independent variable to the prediction is clearly encoded.  
• Flexible predictor functions can uncover hidden patterns in the data - more flexibility because the relationships between independent and dependent variables are not assumed to be linear.  
• Regularization of predictor functions helps fight overfitting- No priori what type of predictive functions is needed. From an estimation standpoint the use of regularized nonparametric functions avoids the pitfalls of dealing with higher order polynomial terms in linear models. From an accuracy standpoint, GAM is competitive with popular learning techniques. | • Model results - May produce strange results  
• Forest Plot limitation - Much like a black box and smoothness cannot be controlled of the predictor functions.  
• Bias - Unable to combat bias variance tradeoff as directly as with GAM or ensure interpretable predictor functions. |

### 4.0 APPLICATION IN PRIMARY CARE

#### 4.1 Stochastic Model (Frontier approaches)

Cost efficiency in primary care contracting is obtained when the cost of providing a given level of quantity and quality of primary care services to a specific population (the output contracted by the purchaser) is minimised. Thus, an efficient purchasing agency will buy a given output from a primary care provider incurring the minimum cost, given the underlying production technology in the provision of services. Economic theory can easily predict that in the presence of a sufficient degree of purchaser and provider competition the price paid for a given output would tend to be that corresponding to the minimum cost of production, given the technology and the input prices. Consider primary care provider i contracting services with a
public purchaser over period $t$, and employing $k$ inputs $\mathbf{x}_i = (x_1; \ldots ;x_k)^T$ available at fixed prices $\mathbf{w}_i = (w_1; \ldots ;w_k)^T > 0$ to produce $n$ outputs $\mathbf{y}_i = (y_1; \ldots ;y_n)^T$ representing primary health-care services for a given population area (basic health area $i$, given that only one provider is contracted for each area). If the provider and the purchaser behave competitively, the purchasing cost function is represented by $C(\mathbf{y}_i, \mathbf{w}_i)$, which represents the minimum cost of producing output $\mathbf{y}$ ($n$ primary care services provided to the population of area $i$) at input prices $\mathbf{w}$ under normal production conditions (Puig, 2004).

In practice, the actual cost of purchasing primary care services for the population in area may differ from the efficient cost for several reasons. One reason is that the actual cost of purchasing certain services may be either either higher or lower than the minimum cost as a result of favourable or unfavourable unsystematic influences (random shocks). In this case, market structure and regulatory constraints may prevent the minimisation of the cost of purchasing any given level of output and the public authority may fail to achieve cost efficiency in purchasing services for population area $i$. In this case, the actual cost lies above the cost frontier (expected minimum level).

A study done by Lordan (2007) to estimate efficiencies for ‘out-of-hours’ (OOH) cooperatives of primary care services in Northern Ireland (NI) and Republic of Ireland (ROI). The study showed specification changes considered relate to distribution choice for the one sided error, functional form choice and placement of case-mix variables in the model. This analysis used an approach SPF where payroll was modelled as the OOH coops output and services were seen as the inputs that produced payroll. It was argued that these services are exogenous. Results from the analysis suggest that functional form choice does not greatly affect efficiency estimates except in the presence of multi-collinearity. The findings suggest that ROI OOH coops have efficiency gains over their NI colleagues and nurse triaging has efficiency gains over doctor triaging.

A study was done on average purchasing cost efficiencies of primary care contracts in Catalonia using Cobb-Douglas stochastic frontier model showed that purchasing contracts were more cost efficient in the case of Primary Health-care Team’s (PHCT’s) located in urban compared to rural. However, when list size increases, which is more likely to happen in urban areas, a seemingly unjustified diseconomy of scale appears, as purchasing cost increases. The increase in inefficiency associated with the list size of the population served by a PHCT indicates an unjustified higher purchasing cost per unit of output when the PHCT serves a more highly populated area. Another findings were the ratio of nurses per doctor shows that the purchasing cost in those PHCT’s with a higher ratio of nurses is not efficient. The shortage in the number of nurses per doctor has been taken as an indication of allocative inefficiency in the production of health care services in Spain. This result could be attributed to the fact that nurses’ time is not used as a substitute for doctors’ time. Instead, nurses’ time would be devoted to their own activities (diet counselling, obesity control, etc.), which are either not very efficient or are not well captured in our model (Puig-Junoy, 2004).

These studies showed that SFA was able to analyse for efficiency of services in primary care setting but it has few limitation such as multi-co linearity and when the sizes of list (factors in the modelling) increase. This could make the model become unjustified and not well captured due to increase in DMU’s.
4.2 Data-envelopment Analysis (DEA Model)

Over the past decades, DEA has increasingly been applied within the primary care (PC) context showing its suitability for this setting. DEA has a number of features which make it an attractive tool for efficiency measurement of PC delivery: it can handle effectively the existence of multiple PC resources (hereinafter referred as inputs) and multiple health outcomes (hereinafter referred as outputs) in the transformation process. Furthermore, it does not require strong assumptions about the underlying technology linking the inputs to the outputs, and it measures efficiency in relative instead of absolute terms (Amoîdo & Dyson, 2008).

Efficiency has been assessed in terms of different concepts including technical, scale, and cost efficiency. All of the included DEA applications were focused on technical efficiency –i.e. producing the maximum amount of output from a given amount of input, or alternatively producing a given output with minimum quantities; in a number of DEA applications the main focus was the extent to which evaluated organizations could take advantage of returns to scale by altering its size towards optimal scale of delivery (Pelone et al., 2015).

Data Envelopment Analysis is rooted in the concept of efficiency. In general, this term reflects the degree of success with which an organisation uses its inputs/resources x to produce outputs y of a given quality. This can be assessed in different categories of efficiency. Technical efficiency is determined by comparing the difference between the observed ratio of combined quantities of an organization's output to input and the ratio achieved by best practice. Producing the maximum output or consuming the minimum inputs, as compared to what is technically feasible, is an essential step for service providers to be able to meet their objectives best. According to this concept an organisation or facility (decision-making unit (DMU)) is efficient, if it operates on its corresponding production possibilities frontier. Inefficient producers operate below it. Generally, the best practice frontier can be determined either by using parametric approaches, which are based on regression analysis, or by applying nonparametric techniques. Data Envelopment which was proposed by Charnes et al. (1962) in its present form, is linked with the concept of relative efficiency of similar entities. First, DEA does not impose the assumption of any functional form on the relationship between inputs and outputs. This attribute is especially useful for cases in which the correspondence is not known or specified by theory.

DEA uses linear programming to construct a piecewise efficiency frontier and must only be based on the minimum assumptions of monotonicity and convexity of the efficiency frontier. Second, DEA can be used not only to identify inefficient units, but also to estimate the degree of inefficiency. Third, it is possible to include multiple inputs and outputs in DEA, which is especially important for the analysis of health care services, whereas the weights are calculated within the DEA procedure. The DEA method allows variable and fixed inputs, whereas the variable inputs might change in the short run while the values of the fixed inputs are only allowed to be changed in the long run.

The DEA approach allows each DMU to choose the optimum weightings for the outputs and inputs. Each DMU is considered in turn and its most favourable weights are selected. Special features provide the analysis' adjustment to the concrete problem setting. However, the strengths of DEA lay the foundation of its weaknesses as well. As DEA is an empirically based estimation technique, it is sensible to outliers, error measurements and
random influences in the data. DEA deems any deviation from the efficiency frontier to be the result of inefficiency. From the endogenous weighting system follows a second shortcoming. If the number of factors considered in the efficiency analysis is relatively high, the DEA approach may lead to substantial overestimates of the efficiency of DMUs. Nevertheless, DEA is probably because of its advantages the most appropriate technique currently available for measuring relative efficiency in health services.

According to DEA the efficiency of a multiple output, multiple input DMU \( k \), with \( k = 1, \ldots, n \), can be represented as follows:

\[
\theta = \frac{1}{n} \sum_{j=1}^{s} u_j y_j + \sum_{i=1}^{m} v_i x_i
\]

where \( u \) measures the weight of each output \( y_j \) \( (j = 1, \ldots, s) \), and \( v \) indicates the weight of each input \( x_i \). The efficient frontier of the group (\( \theta = 1 \)) is constituted by the most efficient DMUs to which the efficiencies of the remaining DMUs are related to. Best performers do not waste any input and can therefore be regarded as "peers" for entities with a weak evaluation and a less efficiency. This implies that the efficiency score \( \theta \) falls between 0 and 1.

To estimate the efficiency frontier, different important options have to be fixed. The orientation of the models reflects the appropriate direction of optimisation. This reflects which kind of quantities managers have better under control. In some branches, the organisations may be given a fixed quantity of resources and asked to produce as much output as possible. In this case the output oriented approach might be appropriate. By contrast, an input oriented approach should be used, if a fixed level of output has to be reached by using a minimal quantity of inputs. Because DMUs might differ according to their size and the quantities of inputs used and outputs produced respectively, assumptions concerning the returns to scale have to be formulated.

The Charnes, Cooper, Rhodes (CCR) model incorporates constant returns to scale in production. The efficiency measure (1) and the usual side conditions can be adjusted accordingly. To obtain linear programming problem, the Charnes Cooper Transformation can be used. Based on the CCR approach several other models were developed which build a profound basis for efficiency analysis with different returns to scale, different envelopment surfaces and different ways to project inefficient entities to the efficient frontier. Banker et al. formulated the Banker, Charnes, Cooper (2009) (BCC) model which evaluates solutions for non-increasing returns to scale, non-decreasing returns to scale or variable returns to scale. Whereas the CCR model only measures overall technical efficiency the BCC model exclusively evaluates pure technical efficiency, because scale effects are taken into account.

The comparison of CCR and BCC results enables to identify inefficiency which can be mitigated by increasing or decreasing the production volume resulting in a removal of scale inefficiencies. Thus, the ratio of the CCR and BCC efficiency measures will yield an estimate of the pure scale efficiency (SE) of DMU, example indicating whether the different primary care facilities were operating on an efficient scale in producing their services. The optimal size of a DMU is reached when a marginal increase of all inputs (scale) leads to the same relative increase of outputs. The bigger the difference between the scale efficiency score of a DMU and full scale efficiency (\( SE_k = 1 \)), the more unfavourable are the consequences of scale. It tells us how much output of a DMU can be expanded until it is as efficient as the reference unit (Martic, 2009).

A DEA study done at Burkina Faso on efficiency on primary care suggested that inefficiency is mainly a result of poor utilization of health care facilities as they were either too big or the
demand was too low. The regression results showed that distance was an important factor influencing the efficiency of a health care institution. The further the distance the less efficient the health care institution to the villagers. This study also showed higher in share of relatively shared units. The squared correlation between the observed and predicted efficiency score (McKelvey & Zavaino’s pseudo R²) is 0.219, indicating that the included predictors account for nearly 22% of the variability in the outcome variable; which is rather weak. The value of the ancillary statistic "sigma" (0.296) can be compared with the standard deviation of the independent variable which is 0.254. The difference might be explained by a rather homogenous structure of the primary care facilities in the Burkina Faso sample. The study also indicates that improving the accessibility of primary care facilities will have a major impact on the efficiency of these institutions. Thus, health decision-makers are called to overcome the demandside barriers in accessing health care (Marshall & Flessa, 2011). This study used two stages of DEA to reduce bias in overestimating the efficiency of DMUs.

A study using DEA done in Ontario showed from the raw data on the qualitative indicators of output, Community Health Centres (CHC) practices perform reasonably well. On average, they achieve scores higher than the other three models for three of the eight performance indicators: health promotion, chronic disease management, and comprehensiveness and fare relatively well on prevention and access. The Health Service Organization (HSO) model has the highest average for the continuity of care and access to primary care services variables. Family Health Networks (FHNs) are the best in terms of preventive services. However, once costs are added to the mix in the DEA analysis, CHCs are the least efficient practice sites virtually across the board, whereas the fee-for-service (FFS) model performs the best. The efficiency scores of both the HSO and FHN sites are more evenly distributed across all quartiles in comparison to FFS and CHCs. The data show that these efficiency score rankings are driven by the costs of running the practice. A number of reasons explain the poor efficiency scores of the CHCs. The link between performance indicators and costs may be non-linear, and it therefore may be relatively 16 inexpensive to achieve a low level of performance, but very costly to push these indicators beyond any given threshold. The study showed fee-for-service practices fare so well in the DEA analysis. This is because the broader costs of the FFS approach are borne outside of the practices themselves – costs such as the reputed over-use of specialists. Nevertheless, FFS physicians clearly face incentives to see as many patients as possible given that their remuneration depends upon the number of visits conducted per period of time, while, at the same time, they would want to minimize the costs of running their individual practices because physicians themselves are the residual claimants to the proceeds of the practice. The result showed that practice type matters in evaluating efficiency. How practices are organized and how physicians are remunerated affect the costs associated with providing patient care (Miliken et al., 2011).

Both studies used DEA due to multi DMUs used. DEA are capable of handling multiple inputs and outputs such as in both studies. It also able to analysed and quantified inefficiency such as study done in Burkina Faso where inefficiency is mainly a result of poor utilization of health care facilities. The study in Ontario were able to compare and contrasts four different models of primary care delivery, which model is the most efficient and under what circumstances and its ability to incorporate a variety of inputs and outputs and then weigh them in order to present each practice in the best possible way.
4.3 Cobb-Douglas Model

If taking Cobb-Douglas production function, we get model $\alpha \beta \mu Y = ALK e$. Taking logarithm linearization both sides, we get $\ln Y = \ln A + \alpha \ln L + \beta \ln K + \mu$ (3). $Y$ stands for total output, $A$ is a Technical Efficiency (TE), $L$ is the number of input labour, $K$ is the capital generally referring to capital investment, $\alpha$ is labour output elastic coefficient, $\beta$ is capital output elastic coefficient, $\mu$ is random disturbance. From this model, the main factors are labour resource investment, capital investment and integrated technology level (including the management level, the quality of the labour resource, the introduction of advanced technology). According to the $\alpha$ and $\beta$ combination condition, C-D production function has three types: (1) $\alpha + \beta > 1$, known as increasing returns, it is advantageous to increase output according to the existing technology with the expansion of production scale; (2) $\alpha + \beta < 1$, called diminishing returns, it is the loss outweighs the gain according to the existing technology expanding the production scale to increase the output; (3) $\alpha + \beta = 1$, known as the constant return type, show that the production efficiency and not with the expansion of production scale and improve, only to improve the technical level can raise economic efficiency.

In this study the model involving the main variables are input indicators and output indicators. Input indicators include capital investment and labour. The labour input includes the number of medical staffs. The capital investment includes public expenditure, subsidies for community public health expenditure and prepaid medical insurance expense. Chinese Community Health Cares (CHCs) have six main functions by providing basic clinical services, prevention, health education, women and children’s care, elderly care, immunizations and physical rehabilitation. As noted above, the output variables reflecting CHCs outcome are underlined as follows:

a) Numbers of outpatient service  
b) Numbers of inpatient service  
c) Numbers of management of chronic diseases  
d) Numbers of preventive health care  
e) Numbers of community health education  
f) Numbers of service for childhood immunization  
g) Numbers of service for pre-parental instruction

Primary health service activities are different from the service technical content, labour intensity and the service risk. For instance, the workload between outpatient and inpatient has a significant difference. To accurately calculating the amount of health services, the standard health service must applied. One standard health service is that a qualified doctor or assistant doctor provides a fifteen minutes and satisfied service for a patient. According to Peng Ying-chun (2011) calculation method, we get the catalogue of health services workload, such as outpatient has one service unit and out-call emergency has two service unit and so on. Actually CHCs activities are divided two class, one is medical health services another is public health services (Zhang et al., 2013).

From the function, the study derived the coefficient $\alpha + \beta = 0.596 + 0.239 = 0.835$. Cobb-Douglas production function types due to whether $\alpha + \beta$ equal one or not. Hence, Wald method is used to test the hypothesis: $H_0 : \alpha + \beta \geq 1$  $H_1 : \alpha + \beta < 1$. The statistical results show
that \( F = 159.3866 \), \( P = 0.000 \), did not meet the 0.05 criteria for significance, null hypothesis rejected, indicating that the economic characteristic of Shanghai primary healthcare services was decreasing returns to scale. Production function shows that labour is the dominant factor for output. This is mainly because the medical health services are a labour-intensive service and extensive professional knowledge is critical. Therefore, investment in medical human resources is an appropriate way to improve Shanghai community health service (SCHS) output. This means under present situation when labour input and capital investment and to extend the one time, and output is less than one time. It also implies the SCHS’ efficiency of health service is low. Under present situation, output is shrinking with increasing labour and capital input. Invest on labour and capital cannot improve SCHS efficiency. It will not increase by expanding the scale except for by improving the technology efficiency (TE). In order to improve (TE), the health administration department should enhance medical health interior management level, by constructing institution mechanism and appropriate operation model. At present China has explored family practice systems and innovate the service contents so as to improve efficiency of primary healthcare service (Zhang, 2013).

The Cobb-Doughlas model was not popular in assessing efficiency of health-care services. If used it is always in combination with other model such as SFA. This could be due to its disadvantages such as function assumes perfect competition in the factor market which is unrealistic when talk about health-care. In health-care we did not assume perfect competition. Furthermore it only take two factors a time and neglects other inputs. Health-care services took all factors seriously and analysed it accordingly.

### 4.4 Trans Log Model

A model used here is a typical production analysis model by help of which it is possible to calculate the outcome of the real process, income distribution process and production process. The starting point is a profitability calculation using surplus value as a criterion of profitability. The surplus value calculation is the only valid measure for understanding the connection between profitability and productivity or understanding the connection between real process and production process. A valid measurement of total productivity necessitates considering all production inputs, and the surplus value calculation is the only calculation to conform to the requirement. If we omit an input in productivity or income accounting, this means that the omitted input can be used unlimitedly in production without any cost impact on accounting results.

In primary care we are able to utilize it to estimate and calculate the profitability using the surplus value and understand the transcending points between profitability and productivity. For example, the productivity of the staff and the human resources combined with the machinery and devices present can be applied to the translog model and thus the profitability can be calculated and estimated.
4.5 General Additive Model

4.5.1 Interpretability

When a regression model is additive, the interpretation of the marginal impact of a single variable (the partial derivative) does not depend on the values of the other variables in the model. Hence, by simply looking at the output of the model, we can make simple statements about the effects of the predictive variables that make sense to a nontechnical person. For example, for the graphic illustration above, we can say that the (transformed) expected value of $Y$ increases linearly as $x^2$ increases, holding everything else constant.

In addition, an important feature of GAM is the ability to control the smoothness of the predictor functions. With GAMs, you can avoid wiggly, nonsensical predictor functions by simply adjusting the level of smoothness. In other words, we can impose the prior belief that predictive relationships are inherently smooth in nature, even though the dataset at hand may suggest a more noisy relationship. This plays an important role in model interpretation as well as in the believability of the results.

In primary care we are able to use the interpretability ease of the GAM for the use of measuring of the single variable. In primary care as the studies at times involve single variables only, we are able to fully utilize the GAM to make it easier for the researcher to interpret the results. The statements about the effects of predictive variables that make it easy for the understanding of the policy makers or the decision makers as the results are simple to be read out.

4.5.2 Flexibility and Automation

GAM can capture common nonlinear patterns that a classic linear model would miss. These patterns range from “hockey sticks” – which occur when you observe a sharp change in the response variable – to various types of “mountain shaped” curves:

When fitting parametric regression models, these types of nonlinear effects are typically captured through binning or polynomials. This leads to clumsy model formulations with many correlated terms and counterintuitive results. Moreover, selecting the best model involves constructing a multitude of transformations, followed by a search algorithm to select the best option for each predictor – a potentially greedy step that can easily go awry. There is no problem with GAM. Predictor functions are automatically derived during model estimation. We don’t have to know up front what type of functions we will need. This will not only save us time, but will also help us find patterns we may have missed with a parametric model. Obviously, it is entirely possible that we can find parametric functions that look like the relationships extracted by GAM. But the work to get there is tedious, and we do not have 20/20 hindsight prior to model estimation. The model estimation can be fully utilized in primary care and the model estimation will be beneficial for the researcher. As automation is present the work may be tedious but once implemented the relationships that can be extracted will be helpful to predict the functions that would be needed.
5.0 CONCLUSION

Assessing primary care services have many ways of using economic models, DEA is one of the famous method used due to its simplicity in mathematical form for the production function. DEA also capable of handling multiple inputs and outputs from primary care services and capable of being used with any input-output measurement. Another factor is DEA able to analysed inefficiency for every evaluated DMUs. It is not possible to use Cobb-Douglas model alone to assess efficiency in primary care services. Cobb-Douglas model are always combine with SFA to assess efficiency in primary care. This is due to Cobb-Douglas model need perfect assumption in factor market which is impossible in assessing health care market. SFA is another type of economic modelling used in assessing efficiency in primary care services but difficult to justified the model if DMUs are too large. The use of the trans-log can be a vital tool in primary care as it allows the full estimation and comparison between profitability and productivity. It thus explores the use of production functions and calculates all necessities to better trim the primary care functions and allocate resources accordingly. The use of GAM in research has not been widespread but may be an important tool for the use of single predictors. It is thus encouraged that future public health research include GAM in the research for the ease of interpretation of the results.

REFERENCES


