PREDICTIVE EFFICIENCY ANALYSIS:
A STUDY OF U.S. HOSPITALS

Andrew L. Johnson
Department of Industrial and Systems Engineering, Texas A&M University,
College Station, TX 77840, USA
Graduate School of Information Science and Technology, Osaka University,
Osaka 565-0871, Japan
ajohnson@tamu.edu

Chia-Yen Lee
Institute of Manufacturing Information and Systems, National Cheng Kung University,
Tainan City 701, Taiwan
cylee@mail.ncku.edu.tw

Abstract: Healthcare costs are higher in the U.S. then anywhere else in the world. A significant portion of the costs are generated in hospitals. We investigate both the efficiency and the effectiveness of U.S. community hospitals using the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project 2009-2011 Nationwide Inpatient Sample, a data set which contains all discharges from an approximate 20% sample of hospitals. Here efficiency is the productivity of the hospital measured relative to the most productive hospitals and effectiveness is how closely the hospital produced relative to the forecasted services needed. We find the effectiveness levels are slightly higher than the efficiency levels in both 2010 and 2011 indicating that hospitals are producing closer to the forecasted level than the actual service level needed. Further, both efficiency and effectiveness levels are low indicating a large variability in the level of resources hospitals use to provide the same set of services. The low effectiveness scores indicate that many hospitals have a high level of resources even relative to the forecasted demand providing some evidence for a medical arms race.

Keyword: Proactive DEA; Multiple output cost function; Performance Measurement

1. INTRODUCTION

In 2012, the United States' expenditures on health accounted for 16.9% of GPD, which is 7.5 percentiles points above the OECD average for the same year [1]. Thirty-one percent of U.S. healthcare expenditures are spent solely on hospital care or approximately 5% of GDP [2]. Estimates of the excess cost in the system consistently exceed $750 billion and range as high as half of all healthcare expenditures [3]. These estimates motivate use to quantify the efficiency in hospitals. Because hospitals make-up such a large portion of healthcare expenditures, hospitals are a potential large source of cost savings.

Cost-control and cost-efficiency analyses are familiar to the hospital industry, where concerns over rising costs have been present since the 1950's and 60's [4-6]. It has been more than 25 years since accountability and assessment were hailed as the next revolution in medical care [7]. Valdez et al. [8] emphasize the role potential operational improvements and improved efficiency can make in cost savings. Yet the best models for efficiency measurement in hospitals suffer serious limitations and are rarely applied in practice.

Existing methods analyzing efficiency of hospitals (for a review see Rosko and Mutter [9]) primarily rely on standard applications of data envelopment analysis (DEA) or stochastic frontier analysis (SFA). A particular limitation of these methodologies is that they assume hospitals will be able to perfectly predict customer demands for hospitals services or that hospitals can adjust input resources without any time delays. Based on this assumption, these methods do not attempt to separate the quality of the forecasted for hospital services from the operational performance of the hospital [10]. Therefore, when a hospital is found to be inefficient, the analysis does not provide insight if that inefficiently is coming from a poor forecast or if inefficiency is the result of poor operational performance.

We build on the insights of Lee and Johnson [11-12] who define an effectiveness measure which complements the efficiency measure. Here, *effective input* is defined as the optimal input resource used in the production system that generates expected outputs determined by the forecast demand. Furthermore, for effectiveness measure, we use the input-truncated production function, defined as the minimum inputs for resources used in a hospital given the quantities of the expected outputs generated. A hospital is achieving *effective production* if its input level is equal to the effective input level identified by the input-truncated production function is employed.

A low effectiveness measure implies the hospital used more inputs in a particular year than can be justified by efficient operations and forecasted growth for the industry. Persistent low effectiveness would indicate the hospital is expanding resources faster than the forecasted demand is expanding, consistent with a medical arms race.

2. MODELING

In a typical productivity study, we estimate the efficiency via a production function which defines the maximum outputs that a firm or production system can produce given input resources. Let $\boldsymbol{x}$ be a vector of input variable quantifying the input resources, y be the single-output variable generated from production system, and $y^{PF} = f(\boldsymbol{x})$ represent maximal output level given inputs. Consider a multiple-input and multiple-output production process. Let $\boldsymbol{x} \in \mathbb{R}_+^{|I|}$ denote a vector of input variables and $\boldsymbol{y} \in \mathbb{R}_+^{|J|}$ denote a vector of output variables for a production system. The production possibility set (PPS) T is defined as $T = \{(\boldsymbol{x}, \boldsymbol{y}): \boldsymbol{x} \text{ can produce } \boldsymbol{y}\}$. Let $i \in I$ be the input index, $j \in J$ be the output index, and $k \in K$ be the firm index. X_{ik} is the data of the i^{th} input resource, Y_{jk} is the amount of the j^{th} production output, and λ_k is the multiplier for the k^{th} firm. Thus, PPS can be estimated by a piece-wise linear convex function enveloping all observations shown in model (1)

$$\tilde{T} = \{(\boldsymbol{x}, \boldsymbol{y}) | \sum_k \lambda_k Y_{jk} \geq y_j, \forall j; \sum_k \lambda_k X_{ik} \leq x_i, \forall i; \sum_k \lambda_k = 1; \lambda_k \geq 0, \forall k\} \tag{1}$$

Then, efficiency, θ, can be measured using the variable-returns-to-scale (VRS) DEA estimator. Input-oriented technical efficiency is defined as the distance function $D_I(\boldsymbol{x}, \boldsymbol{y}) = \inf\{\theta | (\theta\boldsymbol{x}, \boldsymbol{y}) \in \tilde{T}\}$. If $\theta = 1$, then the firm is efficient; otherwise it is inefficient when $\theta < 1$.

To separate the effects of forecasting from operational performance we will need to make some assumptions about timing. Specifically we will assume a hospital manager knows the production function from period $t-1$ and the forecast for growth in services required when they determine the input levels for period t. Thus, our timing assumptions eliminate the concern of endogeneity that are common in the econometrics literature. Related to this issue we have assumed that all inputs are adjustable once a year, but after the level of inputs has been selected at the beginning of the year, the input levels are held fixed

Input-truncated production function is defined based on the input demand function which transforms the expected output to input level in current period. To maintain generality, expected outputs are hospital-specific, each firm can have a different forecast demand, and the input-truncated production function is defined as the production function truncated by the optimal inputs used by a specific hospital. Let d^{t+1} be the expected output in period $t+1$. The effective input, $x^{E(t+1)}$, is the inverse of the production function in period t. The $x^{E(t+1)}$ is formulated as equation (2), where $f_t^{-1}(\cdot)$ is the inverse production function with respect to period t.

$$x^{E(t+1)} = f_t^{-1}(d^{t+1}) = D_I(\boldsymbol{x}, \boldsymbol{d}^{t+1})\boldsymbol{x} \qquad (2)$$

Figure 1 illustrates the effective input for a single-input and a single-output case. For an observation, firm A, the effective input $X_A^{E(t+1)}$ is calculated by the production function $f_t(\cdot)$ and its expected output level d_A^{t+1} in period $t+1$.

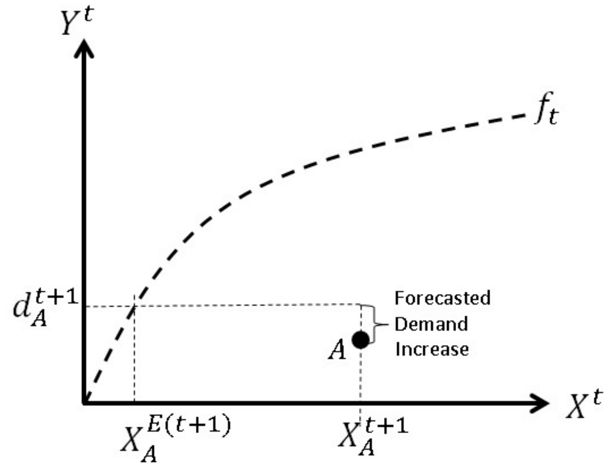

Figure 1 Effective input $x^{E(t+1)}$

To measure the effectiveness, let $\boldsymbol{x}^E \in \mathbb{R}_+^J$ denote an effective input vector estimated *from previous period*. The input-truncated production possibility set (PPS^E)

$$T^E = \{(\max(\boldsymbol{x}^E, \boldsymbol{x}), \boldsymbol{y}): \max(\boldsymbol{x}^E, \boldsymbol{x}) \text{ can produce } \boldsymbol{y} \text{ in current period}\}$$

can be estimated by a piece-wise linear concave function truncated by the effective input level as shown in (3).

$$\tilde{T}^E = \{(\boldsymbol{x}, \boldsymbol{y}) | \sum_k \lambda_k Y_{jk} \geq y_j, \forall j; \sum_k \lambda_k X_{ik} \leq x_i, \forall i; X_i^E \leq x_i, \forall i; \sum_k \lambda_k = 1; \lambda_k \geq 0, \forall k\} \qquad (3)$$

Then, effectiveness, θ^E, can be measured by distance function $D_I(\boldsymbol{x}, \boldsymbol{y}) = \inf\{\theta^E | (\theta^E \boldsymbol{x}, \boldsymbol{y}) \in \tilde{T}^E\}$. If $\theta^E \geq 1$, then the firm is effective in using input resource; otherwise it is ineffective when $\theta^E < 1$ as illustrated in Figure 2.

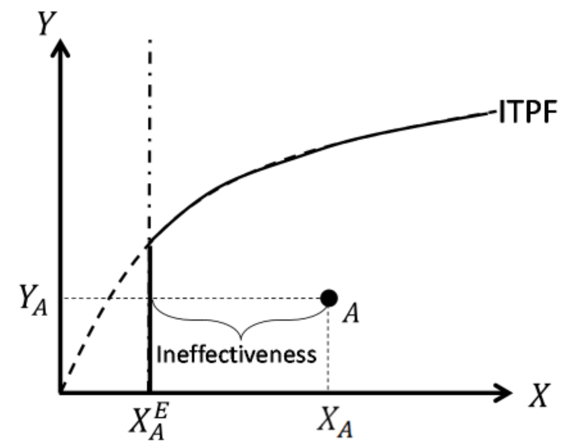

Figure 2 Effectiveness measure

3. RESULTS

In order to examine the effectiveness measure, we use the Agency for Healthcare Research and Quality's (AHRQ) Healthcare Cost and Utilization Project (HCUP) 2009-2011 Nationwide Inpatient Sample, a data set which contains all discharges from an approximate 20% sample

(1,056 hospitals) of U.S. community hospitals as defined by the American Hospital Association. The number of discharges is a single input. We follow [13, 14] and model outputs using a four dimensional vector including: minor diagnostic procedures (y_1), major diagnostic procedures (y_2), minor therapeutic procedures (y_3), and major therapeutic procedures (y_4), categorized by International Classification of Diseases, Clinical Modification codes. The distinguishing characteristic between minor and major procedures of each type is the use of an operating room. For example, an irrigate ventricular shunt is a minor therapeutic procedure, whereas an aorta-renal bypass is a major therapeutic procedure; a CT scan is a minor diagnostic procedure, whereas a brain biopsy is a major diagnostic procedure. In addition, we collect Centers for Medicare and Medicaid Services (CMS) reports which give future projections regarding National Health Expenditure Projections specifically. For example, in 2009, they predict the future industry hospital costs for 2010-2020 and in 2010 they predict 2011-2021 and so forth. We use the expenditure projection to generate the expected output. That is, we take the distribution of outputs from 2009 and multiplied by the expenditure grow projection in 2010 and we have a distribution of the expected 2010 output.

To measure the effectiveness, we select the input level (proxied by the number of discharges) optimally (i.e., $\boldsymbol{x}^E$) given the expected 2010 output with respect to the 2009 frontier. Then we consider the observed outputs and actual discharges for 2010. We use all the data from 2010 to construct a frontier and the hospital specific truncation comes from the $\boldsymbol{x}^E$ estimated from the 2009 data and the 2010 projection. We can now calculate effectiveness relative to the input truncated production function. Thus, when the observed number of discharges in a particular year is larger than the forecasted number of inputs (i.e., $\boldsymbol{x}^E$) we have over-usage of input and the effectiveness is less than 1; otherwise, when the observed discharges is less than (or equal to) the forecasted inputs we have ideal resources and effectiveness is larger than (or equal to) 1.

We do this analysis for two adjacent years 2009-2010 and 2010-2011. After the effectiveness measure we can now look back at the differences between the observed outputs and the distribution of the expected outputs and reconsider if it is best to pick the input level that maximizes the expected performance. Note that we do not observe the same hospitals each year due to the 10% sampling in the hospitals each year and we assume that the collected sample is representative and thereby the distribution of effectiveness characterizes the general population of hospitals.

The results of effectiveness and efficiency regarding 2009-2010 are shown as Figure 3 and Figure 4. Because the data set is an unbalanced panel, there are 279 observations in both adjacent years 2009-2010. The average of effectiveness is 0.521 weighted by the observed inputs in 2010, and the average of efficiency is 0.400.

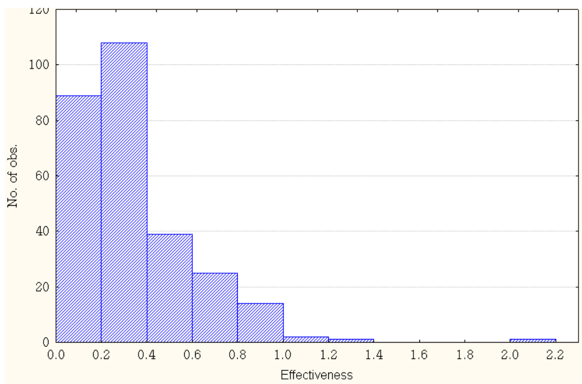

Figure 3 Effectiveness distribution in 2010

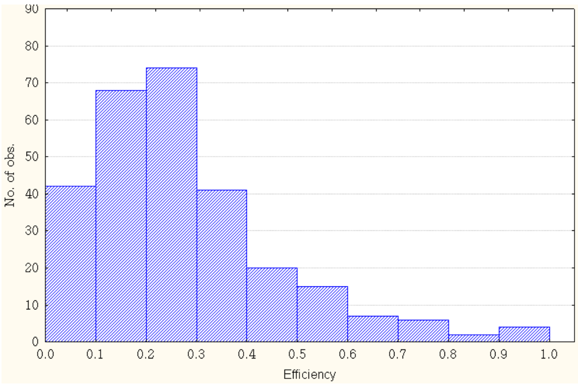

Figure 4 Efficiency distribution in 2010

The results of effectiveness and efficiency for 2010-2011 are shown as Figure 5 and Figure 6, respectively. There are 256 observations present in both adjacent years 2010-2011. The average of effectiveness is 0.504 weighted by the observed inputs in 2010, and the average of efficiency is 0.492.

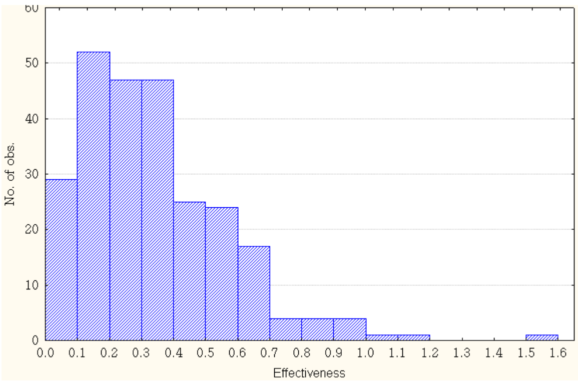

Figure 5 Effectiveness distribution in 2011

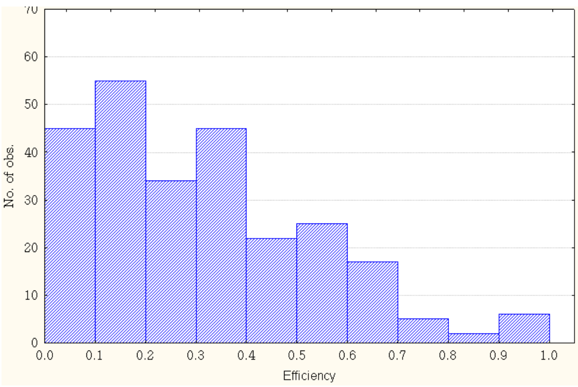

Figure 6 Efficiency distribution in 2011

4. CONCLUSIONS

The efficient operation of hospitals is critical to controlling the costs associated with healthcare in the U.S. An extensive literature exist on measuring efficiency from the inputs consumed and outputs produced by the hospital. For the purposes of evaluating operational performance, this sort of efficiency measure is to combine the effects of forecasting and operational performance. To measure the performance of production units relative to forecasted demand, Lee and Johnson [11] introduced the concept of effectiveness and the truncated production function. We apply these concepts to investigate the performance the U.S. hospital industry.

We find that hospitals measured in terms of efficiency or effectiveness have distributions that are skewed towards having mostly inefficient and ineffective hospitals with a small tall performing relatively well. Having low efficiency and effectiveness scores indicates that it is not primarily differences between the forecast and observed demand that is driving the high inefficiency level results, but appears that operational inefficiency is more systematic. This is in part due to the random nature of demand for hospitals services that requires resources to be available at all times for emergency situations.

In future research we plan to investigate alternative methods for forecasting. In this paper we used the CMS report's National Health Expenditure Projections; however, hospitals within our sample may expect to grow at different rates and therefore use alternative forecasts than CMS. These rates would be driven by local population grow and age.

Using an envelopment estimator such as the Data Envelopment Analysis (DEA) frontier, we find the average efficiency and effectiveness levels are quite low. This may be in part because inefficiency in our model captures noise, inefficiency, and any other unmodeled variables. Therefore, we could use the generalization of

DEA to the stochastic setting that does model noise separate from inefficiency by using a Stochastic Non-parametric Envelopment of Data (StoNED) estimator, see for example [15-17].

REFERENCES

[1] OECD Health Statistics 2014: How does the United States compare? OECD, 2014. [Online] Available at: http://www.oecd.org/unitedstates/Briefing-Note-UNITED-STATES-2014.pdf [Accessed 5 May 2015].

[2] Health Care Cost: A Primer. The Henry J. Kaiser Family Foundation, May 1, 2012. Available at: http://kff.org/report-section/health-care-costs-a-primer-2012-report/ [Accessed 16 September 2015].

[3]The Price of Excess: Identifying Waste in Healthcare Spending. Health Research Institute. PricewaterhouseCoopers, 2009. [Online] Available at: http://www.pwc.com/us/en/healthcare/publications/the-price-of-excess.jhtml [Accessed 16 September 2015].

[4] Sheps, M.C., Approaches to the Quality of Hospital Care, Public Health Reports, 70(9):877-886 (1955).

[5] Dowling, W.L., Hospital production: A linear programming model, Lexington Books, Boston. (1976).

[6] Griffin J.R., Hancock, W.M., and Munson, F.C., Cost Control in Hospitals, Health Administration Press, Ann Arbor. (1976).

[7] Relman, A.S., Assessment and Accountability - The 3rd Revolution in Medical Care, New England Journal of Medicine, 319(18):1220-1222, (1988).

[8] Valdez RS, Ramly E, Brennan PF, Industrial and Systems Engineering and Health Care: Critical Areas of Research--Final Report, AHRQ Publication No. 10-0079. Rockville, MD: Agency for Healthcare Research and Quality. (2010).

[9] Rosko, M.D. and Mutter, R.L., What Have We Learned From the Application of Stochastic Frontier Analysis to U.S. Hospitals?, Medical Care Research and Review, 68(1):75S-100S, (2011).

[10] Lee, C.-Y. and Johnson, A. L.,Two-dimensional efficiency decomposition to measure the demand effect in productivity analysis. European Journal of Operational Research, 216(3), 584–593, (2012).

[11] Lee, C.-Y. and Johnson, A. L., Proactive data envelopment analysis: effective production and capacity expansion in stochastic environments. European Journal of Operational Research, 232(3), 537–548 (2014).

[12] Lee, C.-Y. and Johnson, A. L., Effective production: measuring of the sales effect using data envelopment analysis. Accepted at Annals of Operations Research (2015).

[13] Pope, B. and Johnson, A.L., Returns to Scope: A Metric for Production Synergies Demonstrated for Hospital Production, Journal of Productivity Analysis 40(2): 239-250 (2013).

[14] Sarmento, M., Johnson, A.L., Preciado Arreola, J.L., and Ferrier, G.D., Cost Efficiency of U.S. Hospitals: A Semi-parametric Bayesian Analysis. Working paper.

[15] Kuosmanen, T. and Kortelainen, M., Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints, Journal of Productivity Analysis, 38(1), 11-28, 2012.

[16] Kuosmanen, T. and A.L. Johnson, Data Envelopment Analysis as Nonparametric Least Squares Regression, Operations Research 58(1): 149-160, 2010.

[17] Kuosmanen, T., A.L. Johnson, and A. Saastamoinen, Stochastic nonparametric approach to efficiency analysis: A unified framework, in J. Zhu (Eds) Handbook on Data Envelopment Analysis Vol II, Springer, 2015.