Performance Measurement and Achievable Targets for Public Hospitals

DIEGO PRIOR*
JORDI SURROCA**

This paper proposes an iterative process for determining what we define as a “reasonable frontier efficiency,” a concept that establishes future targets according to their current level of performance. Our “reasonable frontier” has the following properties: (1) it is underpinned on the efficiency framework, (2) it detects the presence of outliers and proposes a procedure to establish the influence introduced by extreme observations, and (3) it sorts out the problem of data masking. The proposed method is applied to a sample of Spanish hospitals. The results obtained confirm the applicability of the proposal made.

Keywords: DEA, Super-efficiency, Data Masking, Target Setting

1. Introduction

A vast body of literature in economics and management is devoted to the problem of measuring the efficiency of hospitals (for an extensive review, see Worthington [2004]). However, the definition of “efficiency” presumes knowledge of the optimal production possibilities. As complete information regarding the transformation function that relates inputs to outputs is generally unknown, parametric approaches may not be suitable. Data Envelopment Analysis hereafter, DEA methods sidestep this problem, and fit a production function using empirical observations from hospitals. A hospital is considered fully efficient if the performance of other units in the sample does not show that some of its inputs or outputs can be improved without worsening some of its other inputs and outputs. Because a set of hospitals is used to estimate the production possibilities, however, a question remains unanswered: How is empirically based...

*Department of Business Economics, Universitat Autònoma de Barcelona, and IESEG School of Management
**Department of Business Administration, Universidad Carlos III de Madrid

We are grateful to Professor Cecilio Mar-Molinero and the anonymous referees for helpful comments and suggestions on earlier drafts of this paper. The authors are also grateful to the Spanish Ministerio de Ciencia e Innovación (grants nos. ECO2010-18967 and ECO2009-08308) and the Dirección General de Universidades de la Comunidad de Madrid (grant no. S2007/HUM-0413) for financial support.
production technology affected by the composition of the sampled hospitals? For example, if the sample includes some superefficient hospitals that have special technological capabilities that cannot be achieved by others, then unachievable targets are likely to be set if DEA is used to estimate potential productivity.

In this paper, we extend the models suggested by Bogetoft (1994) and Agrell, Bogetoft, and Tind (2002) and propose a method for identifying a reasonable production frontier that prescribes achievable targets for hospitals. Previous research proposes estimation methods for identifying superefficient extreme observations. However, the possibility that observations not included in the frontier also may be influential has been ignored. This phenomenon of data masking arises when an extreme observation masks another less extreme but still influential observation; the main contribution of this article is to propose a recursive method to handle such situations. After the process, the inefficient hospitals have as a reference other hospitals that operate more efficiently but with similar environmental characteristics. The advantage of this process is that reliable ranks and achievable targets can be deduced from the assessment of the efficiency analysis.

In related literature, other alternatives have been explored. One of them consists of the use of cluster analysis to classify hospitals according to their characteristics before applying DEA. It is obvious that clustering the hospitals on the basis of their case mix (or their specific input mix) is an effective way to reduce heterogeneity. However, nothing in cluster analysis prevents influential extreme observations from continuing to distort the results in the subsamples.

The structure of this paper is as follows. Section 2 defines the standard DEA and the super-efficiency estimation methods. Section 3 describes our proposed process for building a reasonable frontier. Section 4 illustrates the application of the methodology using a sample of Spanish public hospitals, and includes a discussion of the data, the variables used, the methodology, and the results obtained. Section 5 comments on the managerial relevance of the proposed methodology, as well as its implications and limitations.

2. Performance Valuation Using DEA

In complex organizations such as hospitals, it is difficult to know the exact transformation function that maps input factors into output levels. In these circumstances, benchmarking can be used to assess the performance of inefficient hospitals looking to the achievements presented by more efficient hospitals. DEA is particularly relevant in this process, as it does not attempt to estimate the form of the production function but instead uses existing observations to establish comparisons among all hospitals. From these comparisons, the best performers are identified and used as a reference for the inefficient units.

DEA estimation methods assume the existence of a production possibility set \((T)\) that includes real observations and their convex combinations. It is assumed that
PERFORMANCE MEASUREMENT AND ACHIEVABLE TARGETS

\( T \) satisfies a number of conditions, such as convexity and monotonicity (see Banker & Thrall [1992]), and it is well defined as follows:

\[
T = \left\{ (x, y) : x \geq \sum_{k=1}^{K} \lambda_k x^k, y \leq \sum_{k=1}^{K} \lambda_k y^k, \lambda_k \geq 0, k = 1, \ldots, K \right\}
\]  

(1)

where \( x^k = (x_1^k, \ldots, x_n^k) \in \mathbb{R}^n \) denotes the vector of observed inputs and \( y^k = (y_1^k, \ldots, y_m^k) \in \mathbb{R}^m \) represents the vector of observed outputs. The set \( T \) is limited by an efficient frontier (containing the subsample of the best performers). This frontier can be used to measure the potential output that a hospital can deliver. To do so, it is necessary to solve the following linear programming problem:

Max \( \tilde{\theta}^o \)

s.t. \( \sum_{k=1}^{K} \lambda_k y^k \geq \tilde{\theta}^o y^o_j ; \ j = 1, \ldots, m, \)

\( \sum_{k=1}^{K} \lambda_k x^k \leq x^o_i ; \ i = 1, \ldots, n. \)

(2)

where the superscript \(^o\) refers to the hospital that is under evaluation. The optimal solution to this problem offers an efficiency coefficient, \( \tilde{\theta}^o \geq 1 \), measuring the proportion by which the \( m \) outputs of the hospital being assessed need to increase for the hospital to be located on the production possibility frontier—that is, on the frontier that measures best practices achieved by the most competent hospitals. Under this assessment, a manager is presumed to run a hospital efficiently only when the efficiency coefficient is equal to one. The DEA algorithm identifies a reference set against which the performance of the hospital under observation is assessed. This reference set is associated with nonzero values of the activity vector \( \lambda^* \). In the particular case in which \(^o\) forms part of the frontier, we will find \( \lambda^* = 1 \) and, for the rest of units \( k, \lambda^k = 0 \). The aforementioned information is useful for the assessment of inefficient units—for ranking them according to their efficiency level, for instance; but a problem arises when trying to rank the efficient units, because all of them obtain the same efficiency score (\( \tilde{\theta}^o = 1 \)). To overcome this problem, the use of superefficiency coefficients is suggested; this was initially proposed by Andersen and Petersen (1993). Under the superefficiency estimation, the frontier is defined without including the hospital to be assessed in the reference set. The ordering of hospitals is based on the results of linear programming problem eq. (3), defined as follows:

Max \( \tilde{\theta}^o \)

s.t. \( \sum_{k \neq o}^{K} \hat{\lambda}_k y^k \geq \tilde{\theta}^o y^o_j ; \ j = 1, \ldots, m, \)

\( \sum_{k \neq o}^{K} \hat{\lambda}_k x^k \leq x^o_i ; \ i = 1, \ldots, n. \)

(3)
This new mathematical program implies that the superefficiency coefficient $\hat{\theta}^*$ is no longer constrained to values greater than or equal to unity. In fact, for efficient units, it will provide values smaller than unity, with the values interpreted as the proportion by which hospital $\text{h}_{i}^{*}$ can reduce all outputs and still be considered efficient. This implies that rank-order lists can be established that include the superefficient hospitals.

A second issue may arise when the frontier is defined by hospitals whose performance is difficult to match because they use different technologies or operate in an environment distant from hospital $\text{h}_{i}^{*}$. In the context of frontier analysis, this problem is associated with the presence of extreme and influential observations. In practice, small values of $\hat{\theta}^*$ suggest that we are in the presence of an atypical observation that is far from the frontier.

To identify hospitals whose performance cannot be matched, we can follow the extant literature on identifying influential observations; see, for instance, Timmer (1971), Wilson and Jadlow (1982), Dusansky and Wilson (1994, 1995), Wilson (1995), Pastor, Ruiz, and Sirvent (1999), and Fox, Hill, and Diewert (2004). An influential observation is defined as having a disproportionate impact on the assessment of the rest of the observations (Dusansky & Wilson [1994]), and its detection can be complicated by the existence of multiple outputs (Fox, Hill, & Diewert [2004]), as in the case of public hospitals. One way to identify an influential observation is to remove it from the data set. If the observation is influential, a large change will be seen in the efficiency of the remaining observations. This is not the case when the observation removed is not influential.

In this article, the problem we want to solve is the identification of influential extreme observations when some units do not appear in the original frontier. This will be the case when an extreme observation is masked by a frontier influenced by another more extreme observation, generating a problem of data masking in the frontier efficiency framework. Data masking already has received attention in a parametric context—see, for example, Ezzamel and Mar-Molinero (1990)—but has been ignored in the DEA context. In the following section, we suggest a recursive process for identifying hospitals whose performance cannot be matched.

3. Performance Assessment by Means of a Reasonable Frontier

3.1 The Concept of the Reasonable Frontier

Here we propose an iterative process for efficiency assessment. The first step is the estimation of Andersen and Petersen’s (1993) superefficiencies for all hospitals using linear programming problem eq. (3). Hospitals are ranked in increasing order of superefficiency. The hospital with the lowest value of $\hat{\theta}^*$ is thus identified. Next, we assess the impact that this hospital, $k$, has on the definition of the frontier. Because it has the lowest value of $\hat{\theta}^*$, the exclusion of hospital $k$ from efficiency calculations would produce the largest shift in the position of the frontier. Following this reasoning, we remove hospital $k$ from the sample and recalculate efficiencies. If
the change in efficiency values is significant, we exclude hospital \( k \) from the best practice frontier, as its presence may define targets that are impossible to achieve.

### 3.2 Identification of the Reasonable Frontier

Imagine that, having solved linear programming problem eq. (3), the super-efficient hospital \( r_r \), \( r_2 \); \( :\cdots :K(\) is associated with the lowest value of the efficiency index \( \hat{\theta}_{[r]} \). We need to test whether this hospital should be labeled as influential. To do this, we remove \( r \) from the sample and recalculate the efficiencies for all remaining hospitals by solving the following problem:

\[
\begin{align*}
\text{Max} & \quad \hat{\theta}_{-r}^c \\
\text{s.t.} & \quad \sum_{k=1}^{K} \sum_{j=1}^{m} \hat{z}_k^j y_k^j \geq \hat{\theta}_{-r}^c y_j^o ; \quad j = 1, \ldots, m \\
& \quad \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{z}_k^i x_k^i \leq x_i^o ; \quad i = 1, \ldots, n
\end{align*}
\]

where subindex \(-r\) indicates that observation \( r \) has been left out of the computations. To determine whether \( r \) is an influential observation, we compare super-efficiency coefficients obtained under linear programming problem eq. (3), excluding the one associated with \( r \), \( \hat{\theta}^s = (\hat{\theta}_1^s, \ldots, \hat{\theta}_{-1}^s, \hat{\theta}_{+1}^s, \ldots, \hat{\theta}_K^s) \), and efficiency indexes obtained under linear programming problem eq. (4), \( \hat{\theta}_{-r}^s = (\hat{\theta}_1^s, \ldots, \hat{\theta}_{-1}^s, \hat{\theta}_{+1}^s, \ldots, \hat{\theta}_K^s) \). Afterward, we can say that \( r \) is an influential observation if the distributions of coefficients \( \hat{\theta}^s = (\hat{\theta}_1^s, \ldots, \hat{\theta}_{-1}^s, \hat{\theta}_{+1}^s, \ldots, \hat{\theta}_K^s) \) and \( \hat{\theta}_{-r}^s = (\hat{\theta}_1^s, \ldots, \hat{\theta}_{-1}^s, \hat{\theta}_{+1}^s, \ldots, \hat{\theta}_K^s) \) are statistically different. If statistically significant differences are found, \( r \) is removed from the sample and the procedure is repeated with a new sample that excludes \( r \), \( k = (1, \ldots, r - 1, r + 1, \ldots, K) \). This iterative process ends when the null hypothesis of equality of distributions cannot be rejected.

Diagnosing whether or not an observation is influential obviously depends on the ability to detect statistically significant changes in the distribution of coefficient \( \theta^s \). We thus are assuming that the DEA efficiency coefficient is a random variable with a statistical distribution function, perhaps because the original data contains errors. The \( F \) distribution is a possible statistical distribution for DEA efficiency; see Banker (1993). Several statistics have been used in the DEA

1. Had eq. (3) been input-oriented, then we would have chosen the hospital with the largest efficiency coefficient (superefficient units would have efficiency coefficients greater than one and inefficient units lower than one).
2. The structure of Banker’s analysis depends on specific parametric assumptions concerning the distribution of the inefficiencies (Banker uses either half-normal or exponential distributions). In contrast, such assumptions are not necessary when robust nonparametric statistical procedures are used. Additionally, the asymptotic statistical technique requires large sample sizes, while nonparametric procedures can be used for any sample size.
literature to test hypotheses on the distribution of efficiency. The Wilcoxon-Mann-Whitney rank-sum test has been popular for comparing the frontiers of two different types of units (see, for example, Valdmanis [1992]). Other tests, such as Banker’s (1993) and Kolmogorov-Smirnov also have been used. Kittelsen (1993) argues that, if two measures of efficiency are based on the same units, the correct test to use is the Wilcoxon signed-rank test.

In this particular case, we use the nonparametric Wilcoxon (matched-pairs) signed-rank test for related samples. We chose this test for two reasons. First, we are not obliged to make assumptions about the distribution of efficiency as calculated by \( LP \) eq. (4). Second, we cannot use a test for unrelated samples, such as the Kruskal-Wallis analysis, because the same units belong to both populations.

Wilcoxon’s test checks to see whether two probability distributions share the same median by calculating the following statistic:

\[
z = \frac{T_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{5}\]

where \( n \) is the number of observations in the sample. In the first iteration \( n = I - 1 \), because we are dealing with the complete sample after removing observation \( r \), whose influence we are trying to establish. \( T_+ \) is either the sum of positive ranks or the sum of negative ranks, whichever is lower. The calculation procedure is as follows. We start by calculating the difference between the values of \( \hat{\theta}^r = (\hat{\theta}^{1r}, \ldots, \hat{\theta}^{ir}, \ldots, \hat{\theta}^{Kr}) \) and \( \hat{\theta}^r = (\hat{\theta}^{1r}, \ldots, \hat{\theta}^{Kr}) \). Next, we order the absolute values of the differences. We assign ranks from one to \( n \), with one being associated with the smallest difference and \( n \) with the highest difference. We add the ranks of all the cases associated with positive differences to obtain the sum of positive ranks, and we proceed in the same way with the ranks of negative differences to obtain the sum of negative ranks.

Hospital \( r \) is considered an influential observation if the null hypothesis is rejected. If \( r \) is identified as influential, it should be removed. The final result after the successive removal of all influential observations is what we define as the reasonable frontier. The basic properties of the reasonable frontier are that (1) it is not influenced by the presence of efficient hospitals whose performance it cannot match; (2) it solves the problem of influential observations that are masked by unambiguous outliers; and (3) it is representative of the sample, as it does not change when an additional unit is removed.

To guide the imagination, Figures 1 and 2, referring to a single-input/two-outputs situation, are included. In Figure 1, an extreme and influential observation \( r_1 \) is

---

3. However, if models are nested, this test is not significant, because every change in the efficiency coefficients is the same for the entire sample. Pearson or Spearman correlation coefficients can be applied with nested models.

4. If there are equal differences, the same average rank value is assigned to the units.
presented. Program eq. (3) will detect $r_1$ as a superefficient observation. Moreover, Wilson’s (1995) proposal will characterize $r_1$ as an influential and extreme observation because the original frontier, which includes $r_1$ and is represented by the black dotted line, is distant from the frontier that excludes $r_1$, which is represented by the grey solid line and is closer to the remaining observations.

Figure 2 illustrates this situation. As shown in Figure 1, the Wilson method will detect $r_1$ as extreme and influential. However, the existent methods will never find $r_2$, as this observation does not form part of the subsample of efficient units when applying program eq. (3). Our proposal will detect $r_2$ after the second iteration, which implies that the reasonable frontier will be defined by the grey solid line instead of the excessively demanding grey dotted line that includes $r_2$.

Figure 2 illustrates what can be done when important differences that separate extreme observations from the rest appear. On the one hand, the efficiency regarding the units remaining in the original sample can be assessed with the reasonable frontier (the grey solid line in Figure 2); on the other hand, the removed observations can be assessed by using their specific frontier (the black dotted line in Figure 2). In this way, both subsamples are compared with their respective achievable potential targets. In normal situations, this is far more desirable than the original assessment using linear program eq. (2), which imposes targets impossible to achieve for the majority of the observations.
4. Data and Results

The data used in this article were obtained from the EESRI (Estadística de Establecimientos Sanitarios con Régimen de Internado, Ministerio de Sanidad y Consumo) survey of the Spanish Ministry of Health. The survey contains information on all public hospitals in Spain. This study is limited to general hospitals (specialized hospitals and nursing homes are excluded) with more than 200 beds, thus restricting the number of observations to 137.

Following Chillingerian and Sherman (1990), the concept of efficiency used in this article deals with hospital management efficiency, but the analysis could be extended to the efficiency of medical practitioners in the provision of health services.5 The variables selected to assess efficiency are as follows:

- \( y_1 \): Total number of discharges: The number of discharges from internal medicine, medical specialties, general and digestive surgery, traumatology, obstetrics, and gynecology.
- \( y_2 \): Total number of bed-days: The number of bed-days in internal medicine, medical specialties, general and digestive surgery, traumatology, obstetrics, and gynecology.

5. The EESRI questionnaire does not contain the required information for the evaluation of hospital activity in terms of internal and external medical efficiency (Zweifel, Breyer, & Kifmann [2009]). As pointed out by Murray (1992), if, in variable selection, we are interested in determining the final outcome of health services, there are enough reasons to prefer outcome variables (i.e., the number of patients treated weighted by their corresponding Diagnostic Related Group) rather than throughput variables (i.e., the number of patient days or the number of cases treated).
**PERFORMANCE MEASUREMENT AND ACHIEVABLE TARGETS**

\( y_3: \) Total number of outpatient appointments: Medical assistance to outpatients for diagnostic purposes, for treatment, and for the follow-up of clinical cases.

The following productive resources were considered:

\( x_1: \) Medical staff and other professional workers: Medical staff (full-time equivalent at forty hours per week), including pharmacists, medical staff on hospital duty, and other staff with higher degrees.

\( x_2: \) Other staff: Nurses, physiotherapists, auxiliary staff, managers, and administration staff.

\( x_3: \) Beds: Number of hospital beds as a proxy for investment in plant, and as a measure of the size of hospital, building volume, and plant complexity.

\( x_4: \) External purchases: Purchases of medicines, sanitary material, food and drinks, instruments, clothes, and other consumables.

Table 1 shows the descriptive statistics for the input and output variables. All the inputs and outputs are essential, as positive values appear for all the hospitals in the sample. Panels B and C identify the characteristics corresponding to the extreme observations. Comparing the descriptive statistics presented on these panels, it appears that the extreme units are mainly the smaller hospitals. Smaller hospitals usually operate with a less severe case mix. As we could not obtain information about the severity of the cases for each hospital, it may well be that differences in the severity of the case mix the hospitals are treating separates these influential small hospitals from the rest.

Table 2 summarizes the results of applying our proposed method to the sample of hospitals. The first five columns of the table contain information on the iterative process required for the identification of the reasonable frontier. There is an alternative procedure for outlier identification in DEA developed by Wilson (1995). Wilson’s procedure was also applied: the next five columns of Table 2 contain relevant information on this alternative method. The final three columns of Table 2 compare the results of Wilson’s procedure with that of our proposed method, and indicate the decision that we would take about the removal of each observation from the data set on which the reasonable frontier is based.

The first row of Table 2 deals with Hospital 22 (H22), which was considered for inclusion in the set of extreme and influential observations. Linear programming problem eq. (3) returns an efficiency of 45.92 percent, indicating that this hospital could reduce its outputs to about 46 percent of their current levels and still be efficient. If we look at the complete sample, this is the hospital that is most distant from the frontier on the superefficiency side; as such, it is the principal candidate for extreme observation status. When hospital H22 is removed from the data set and the efficiency of all remaining hospitals is recalculated using linear programming problem eq. (4), the average inefficiency level decreases from 110.66 percent to 110.43 percent. According to Wilcoxon’s test, this is a significant change at the 1 percent probability level. Based on the procedure suggested, H22 should be removed from the sample. A similar conclusion
### TABLE 1

Descriptive Statistics

**Panel A**: Initial sample. \( N = 137 \)

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 ): Medical staff</td>
<td>569.83</td>
<td>361.97</td>
<td>200.00</td>
<td>1995.00</td>
</tr>
<tr>
<td>( x_2 ): Other staff</td>
<td>589.55</td>
<td>427.17</td>
<td>66.34</td>
<td>2,447.96</td>
</tr>
<tr>
<td>( x_3 ): Beds</td>
<td>1,597.57</td>
<td>1,238.75</td>
<td>274.55</td>
<td>6,878.60</td>
</tr>
<tr>
<td>( x_4 ): External purchases</td>
<td>3,100,886.18</td>
<td>2,536,931.12</td>
<td>397,836.00</td>
<td>1,332,2528.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 ): Discharges</td>
<td>21,081.72</td>
<td>13,169.63</td>
<td>3,775.00</td>
<td>67,787.00</td>
</tr>
<tr>
<td>( y_2 ): Bed-days</td>
<td>171,058.50</td>
<td>111,330.92</td>
<td>31,460.00</td>
<td>54,2702.00</td>
</tr>
<tr>
<td>( y_3 ): Outpatients</td>
<td>159,007.64</td>
<td>98,274.04</td>
<td>31,143.00</td>
<td>48,0810.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Efficiency (linear problem eq. (2))</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>112.05%</td>
<td>9.51%</td>
<td>100.00%</td>
<td>146.38%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Efficiency (linear problem eq. (3))</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>110.66%</td>
<td>12.69%</td>
<td>45.92%</td>
<td>146.38%</td>
</tr>
</tbody>
</table>

**Panel B**: Final sample. \( N = 119 \)

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 ): Medical staff</td>
<td>608.54</td>
<td>370.07</td>
<td>201.00</td>
<td>1,995.00</td>
</tr>
<tr>
<td>( x_2 ): Other staff</td>
<td>642.45</td>
<td>431.18</td>
<td>165.30</td>
<td>2,447.96</td>
</tr>
<tr>
<td>( x_3 ): Beds</td>
<td>1,747.61</td>
<td>1,255.14</td>
<td>430.89</td>
<td>6,878.60</td>
</tr>
<tr>
<td>( x_4 ): External purchases</td>
<td>3,399,683.29</td>
<td>2,579,357.12</td>
<td>755,371.00</td>
<td>13,322,528.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y_1 ): Discharges</td>
<td>22,528.40</td>
<td>13,376.00</td>
<td>7,281.00</td>
<td>67,787.00</td>
</tr>
<tr>
<td>( y_2 ): Bed-days</td>
<td>183,040.67</td>
<td>113,534.19</td>
<td>55,526.00</td>
<td>542,702.00</td>
</tr>
</tbody>
</table>


### TABLE 1 (Continued)

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>162,131.62</th>
<th>98,767.35</th>
<th>31,143.00</th>
<th>480,810.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>(linear problem eq. (2))</td>
<td>108.66%</td>
<td>7.21%</td>
<td>100.00%</td>
<td>125.61%</td>
</tr>
<tr>
<td>(linear problem eq. (3))</td>
<td>107.38%</td>
<td>9.12%</td>
<td>88.93%</td>
<td>125.61%</td>
</tr>
</tbody>
</table>

**Panel C: Outliers, N = 18**

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_1$: Medical staff</td>
<td>313.94</td>
<td>133.31</td>
<td>200.00</td>
</tr>
<tr>
<td>$x_2$: Other staff</td>
<td>239.82</td>
<td>142.15</td>
<td>66.34</td>
</tr>
<tr>
<td>$x_3$: Beds</td>
<td>605.58</td>
<td>376.25</td>
<td>274.55</td>
</tr>
<tr>
<td>$x_4$: External purchases</td>
<td>1,125,505.22</td>
<td>742,500.47</td>
<td>397,836.00</td>
</tr>
<tr>
<td><strong>Outputs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_1$: Discharges</td>
<td>11,517.50</td>
<td>5,840.51</td>
<td>3,775.00</td>
</tr>
<tr>
<td>$y_2$: Bed-days</td>
<td>91,843.06</td>
<td>45,116.69</td>
<td>31,460.00</td>
</tr>
<tr>
<td>$y_3$: Outpatients</td>
<td>138,354.61</td>
<td>95,032.82</td>
<td>33,520.00</td>
</tr>
</tbody>
</table>

Source: AQ1

---

**TABLE 1**

<table>
<thead>
<tr>
<th>Efficiency</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_2$: Bed-days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y_3$: Outpatients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 2
Influential Observations

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Potentially influential hospital</th>
<th>Andersen-Petersen (CRS)</th>
<th>Average sample efficiency including $r$</th>
<th>Wilcoxon’s $Z$</th>
<th>Potentially influential hospital</th>
<th>Andersen-Petersen (VRS)</th>
<th>Unit whose efficiency changes when $r$ is removed</th>
<th>Sum of efficiency changes / $n_r$ observations</th>
<th>Total efficiency change</th>
<th>Wilson Ranking</th>
<th>Our procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$H22$</td>
<td>45.92%</td>
<td>11.066%</td>
<td>−3.92***</td>
<td>$H22$</td>
<td>43.14%</td>
<td>17</td>
<td>4.42%</td>
<td>75.15%</td>
<td>Wilson</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>$H20$</td>
<td>57.31%</td>
<td>11.043%</td>
<td>−3.29***</td>
<td>$H20$</td>
<td>n.d.</td>
<td>11</td>
<td>2.10%</td>
<td>29.36%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$H25$</td>
<td>71.29%</td>
<td>11.045%</td>
<td>−2.36**</td>
<td>$H25$</td>
<td>78.26%</td>
<td>11</td>
<td>1.27%</td>
<td>13.94%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$H75$</td>
<td>75.91%</td>
<td>11.073%</td>
<td>−2.36**</td>
<td>$H75$</td>
<td>78.93%</td>
<td>1</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$H98$</td>
<td>80.60%</td>
<td>11.061%</td>
<td>−2.93***</td>
<td>$H98$</td>
<td>72.50%</td>
<td>24</td>
<td>0.84%</td>
<td>20.14%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$H31$</td>
<td>66.99%</td>
<td>11.059%</td>
<td>−4.45***</td>
<td>$H31$</td>
<td>88.07%</td>
<td>16</td>
<td>0.35%</td>
<td>5.53%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$H107$</td>
<td>79.64%</td>
<td>11.066%</td>
<td>−3.70***</td>
<td>$H107$</td>
<td>73.88%</td>
<td>10</td>
<td>1.37%</td>
<td>13.68%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$H104$</td>
<td>78.10%</td>
<td>11.067%</td>
<td>−4.10***</td>
<td>$H104$</td>
<td>84.07%</td>
<td>7</td>
<td>1.21%</td>
<td>8.45%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>$H28$</td>
<td>83.44%</td>
<td>11.073%</td>
<td>−7.10***</td>
<td>$H28$</td>
<td>86.20%</td>
<td>48</td>
<td>1.05%</td>
<td>50.20%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>$H47$</td>
<td>74.87%</td>
<td>11.007%</td>
<td>−9.10***</td>
<td>$H47$</td>
<td>82.46%</td>
<td>57</td>
<td>1.58%</td>
<td>89.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$H19$</td>
<td>83.30%</td>
<td>107.89%</td>
<td>−4.45***</td>
<td>$H19$</td>
<td>99.59%</td>
<td>3</td>
<td>0.15%</td>
<td>0.44%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>$H87$</td>
<td>84.83%</td>
<td>107.73%</td>
<td>−3.35***</td>
<td>$H87$</td>
<td>98.02%</td>
<td>3</td>
<td>0.25%</td>
<td>0.75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>$H23$</td>
<td>84.24%</td>
<td>107.54%</td>
<td>−2.42%</td>
<td>$H23$</td>
<td>98.43%</td>
<td>26</td>
<td>1.30%</td>
<td>33.76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>$H26$</td>
<td>87.20%</td>
<td>107.67%</td>
<td>−4.37***</td>
<td>$H26$</td>
<td>98.02%</td>
<td>3</td>
<td>0.25%</td>
<td>0.75%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>$H50$</td>
<td>87.76%</td>
<td>107.59%</td>
<td>−1.75*</td>
<td>$H50$</td>
<td>n.d.</td>
<td>2</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>$H35$</td>
<td>87.99%</td>
<td>107.71%</td>
<td>−6.14***</td>
<td>$H35$</td>
<td>84.53%</td>
<td>26</td>
<td>1.30%</td>
<td>33.76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>$H3$</td>
<td>88.98%</td>
<td>107.45%</td>
<td>−3.40***</td>
<td>$H3$</td>
<td>84.53%</td>
<td>26</td>
<td>1.30%</td>
<td>33.76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>$H64$</td>
<td>87.73%</td>
<td>107.34%</td>
<td>−2.66***</td>
<td>$H64$</td>
<td>84.53%</td>
<td>26</td>
<td>1.30%</td>
<td>33.76%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>$H82$</td>
<td>88.93%</td>
<td>107.38%</td>
<td>−1.60</td>
<td>$H82$</td>
<td>84.53%</td>
<td>26</td>
<td>1.30%</td>
<td>33.76%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**TABLE 2 (Continued)**

<table>
<thead>
<tr>
<th>H17</th>
<th>99.82%</th>
<th>1</th>
<th>0.00%</th>
<th>0.00%</th>
<th>36</th>
<th>36</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>H27</td>
<td>78.32%</td>
<td>33</td>
<td>0.75%</td>
<td>24.64%</td>
<td>7</td>
<td>9</td>
<td>No</td>
</tr>
<tr>
<td>H33</td>
<td>99.80%</td>
<td>3</td>
<td>0.22%</td>
<td>0.65%</td>
<td>35</td>
<td>29</td>
<td>No</td>
</tr>
<tr>
<td>H30</td>
<td>80.42%</td>
<td>42</td>
<td>1.70%</td>
<td>71.37%</td>
<td>9</td>
<td>3</td>
<td>No</td>
</tr>
<tr>
<td>H118</td>
<td>68.92%</td>
<td>7</td>
<td>1.05%</td>
<td>7.34%</td>
<td>3</td>
<td>18</td>
<td>No</td>
</tr>
<tr>
<td>H95</td>
<td>96.10%</td>
<td>39</td>
<td>0.64%</td>
<td>24.88%</td>
<td>25</td>
<td>8</td>
<td>No</td>
</tr>
<tr>
<td>H51</td>
<td>89.32%</td>
<td>7</td>
<td>0.40%</td>
<td>2.82%</td>
<td>18</td>
<td>26</td>
<td>No</td>
</tr>
<tr>
<td>H119</td>
<td>92.51%</td>
<td>1</td>
<td>0.60%</td>
<td>0.60%</td>
<td>21</td>
<td>30</td>
<td>No</td>
</tr>
<tr>
<td>H37</td>
<td>96.00%</td>
<td>21</td>
<td>0.75%</td>
<td>15.84%</td>
<td>24</td>
<td>13</td>
<td>No</td>
</tr>
<tr>
<td>H18</td>
<td>92.53%</td>
<td>14</td>
<td>0.42%</td>
<td>5.93%</td>
<td>22</td>
<td>21</td>
<td>No</td>
</tr>
<tr>
<td>H16</td>
<td>98.56%</td>
<td>13</td>
<td>0.22%</td>
<td>2.92%</td>
<td>29</td>
<td>25</td>
<td>No</td>
</tr>
<tr>
<td>H21</td>
<td>98.23%</td>
<td>11</td>
<td>0.31%</td>
<td>3.40%</td>
<td>28</td>
<td>24</td>
<td>No</td>
</tr>
<tr>
<td>H24</td>
<td>99.63%</td>
<td>6</td>
<td>0.09%</td>
<td>0.51%</td>
<td>34</td>
<td>31</td>
<td>No</td>
</tr>
<tr>
<td>H46</td>
<td>93.88%</td>
<td>4</td>
<td>0.92%</td>
<td>3.66%</td>
<td>23</td>
<td>23</td>
<td>No</td>
</tr>
<tr>
<td>H48</td>
<td>99.37%</td>
<td>7</td>
<td>0.06%</td>
<td>0.39%</td>
<td>32</td>
<td>33</td>
<td>No</td>
</tr>
<tr>
<td>H68</td>
<td>85.93%</td>
<td>5</td>
<td>4.41%</td>
<td>22.03%</td>
<td>14</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>H74</td>
<td>90.64%</td>
<td>4</td>
<td>1.66%</td>
<td>6.63%</td>
<td>19</td>
<td>19</td>
<td>No</td>
</tr>
<tr>
<td>H110</td>
<td>98.83%</td>
<td>4</td>
<td>0.07%</td>
<td>0.26%</td>
<td>30</td>
<td>34</td>
<td>No</td>
</tr>
<tr>
<td>H111</td>
<td>99.96%</td>
<td>3</td>
<td>0.01%</td>
<td>0.04%</td>
<td>37</td>
<td>35</td>
<td>No</td>
</tr>
<tr>
<td>H113</td>
<td>96.93%</td>
<td>3</td>
<td>0.27%</td>
<td>0.82%</td>
<td>26</td>
<td>27</td>
<td>No</td>
</tr>
<tr>
<td>H114</td>
<td>85.10%</td>
<td>38</td>
<td>0.48%</td>
<td>18.29%</td>
<td>13</td>
<td>12</td>
<td>No</td>
</tr>
<tr>
<td>H121</td>
<td>98.88%</td>
<td>33</td>
<td>0.18%</td>
<td>6.08%</td>
<td>31</td>
<td>20</td>
<td>No</td>
</tr>
<tr>
<td>H122</td>
<td>87.69%</td>
<td>17</td>
<td>1.69%</td>
<td>28.65%</td>
<td>16</td>
<td>7</td>
<td>No</td>
</tr>
<tr>
<td>H133</td>
<td>91.49%</td>
<td>18</td>
<td>0.79%</td>
<td>14.16%</td>
<td>20</td>
<td>14</td>
<td>No</td>
</tr>
</tbody>
</table>

**Source:**

**Note:**

***, **, * denote that the null hypothesis can be rejected at the 1, 5, and 10 percent levels. n.d.: superefficiency values not defined.

$\hat{\alpha}'$ and $\hat{\alpha}'_n$ are the equivalent results using variable returns to scale (VRS) to solutions using eqs. (3) and (4), $\theta'$ and $\theta'_n$. 
would be reached if Wilson’s method were used. Wilson proposes two outlier detection criteria. Under both of them, $H_{22}$ takes second place in the ranking of extreme observations; the decision rule would suggest that $H_{22}$ be classified as an outlier. This is in agreement with our iterative procedure.

Differences between the two methodologies begin to appear after iteration three. Iteration three identifies hospital $H_{75}$ as an outlier under Wilcoxon’s test procedure. Under Wilson’s noniterative procedure, $H_{75}$ would never be removed from the data set. It can be seen in Table 2 that its presence affects the efficiency of a single hospital, and its removal does not change the overall efficiency. For another example, take iteration ten. In this iteration, we see that $H_{19}$ is an influential extreme observation. However, this hospital would never be classified as an outlier under Wilson’s procedure (it would be ranked as the thirty-second or thirty-third most extreme observation).

Data-masking problems appear to crop up after iteration eleven and are also present in iterations twelve, fifteen, and seventeen. The methodology suggested here classifies hospitals $H_{87}$, $H_{23}$, $H_{35}$, and $H_{64}$ as extreme observations, but only after other observations, much more extreme than these, have been removed. This ability to cope with masking is absent from Wilson’s (1995) methodology, because it concentrates on the observations that define the original frontier and does not contemplate the possibility that some observations may be revealed as extreme when other, more extreme observations have been removed from the frontier. Only the methodology presented in this paper can cope with the problem of masked extreme observations.

Table 2 gives examples of hospitals that would be classified as extreme observations under Wilson’s method, but they are not detected as such by our iterative method. Hospitals $H_{27}$ and $H_{30}$ are such cases. This reversal of fortune may be due to the removal of hospitals in previous iterations, which has an impact on the statistical test on whose significance we base outlier detection.

The application of the iterative procedure results in the removal of eighteen hospitals, and in the construction of a reasonable and more homogeneous frontier. Table 3 shows that the reasonable frontier is defined by twenty-six hospitals, about double the number of efficient observations in the first iteration. All inefficiency levels have been reduced in the reasonable frontier: (1) average superefficiency is lower, indicating that the increase in outputs to meet the frontier is now far more achievable; (2) the standard deviation of the efficiency coefficients is lower, as less heterogeneity is found after the extreme observations were removed; and (3) both minimum and maximum levels of efficiency are closer to 100 percent. These results ensure that each efficient hospital has a neighboring point in the frontier, and this ensures that the efficiency of the entire sample does not depend much on the presence or absence of a particular hospital. In other words, what we describe as a reasonable frontier is statistically independent of the presence of a particular efficient hospital, as the removal of any observation does not have a significant impact on the rest of the sample; the role of the hospital removed as a reference point is just taken up by another neighboring efficient hospital.
5. Conclusion

Attempting to measure performance provides managers and policymakers with a heightened awareness of shortcomings in data. If such data deficiencies are cataloged and advertised, the quality of data may be improved, thus enhancing the ability to measure performance. The reliability of the performance measures must be subjected to scrutiny. DEA, like other efficiency measurement techniques, is sensitive to outliers. Outliers can arise from either measurement or reporting errors (which may be deliberate), or can reflect significant efficiencies being achieved by particular hospitals that use different production technologies. In these cases, hospitals should be evaluated to determine whether they could be used for purposes of comparison. This is the subject of the present article, and the reasonable frontier is its final outcome.
A critical issue in this context is the selection of hospitals that define the frontier. If the frontier contains hospitals whose performance cannot be matched, this may create disincentives rather than incentives, because the setting of an unrealistic target may result in the hospital we are trying to motivate giving up the challenge. Some studies are devoted to detecting influential observations, but to this point, no proposals have dealt with the data-masking problem. In practice, the recursive process we propose found (masked) extreme observations that previous proposals did not. After controlling for the impact of these observations, we built the best practice frontier by using representative and homogeneous units. Another interesting feature of our approach is that the inefficiency found remains constant even when a particular hospital is removed, so future conservative targets can be defined to improve efficiency.

The limitation of our proposal is the need to consider the context to which models are applied to assess efficiency. It is true that some inputs are fixed while others are variable, but this fact does not necessarily distort the results, as an output orientation is taken in programs eqs. (2), (3), and (4). Therefore, inefficient hospitals should increase their outputs without requiring more inputs (both fixed and variable). Older, less sophisticated hospitals may have less investment per bed, and this can distort the inefficiency coefficients found. The level of some inputs is decided according to the predicted demand rather than the observed demand, but this means that uncertainty, instead of bad management, can cause inefficiency. Because all these circumstances are crucial, they certainly have a direct impact on the results found and should be considered.

REFERENCES


AUTHOR QUERIES

AQ 1: Please provide source for tables and figures where necessary