Published in final edited form as:

J Appl Stat. 2012 ; 39(12): 2639–2653. doi:10.1080/02664763.2012.724664.

Data Envelopment Analysis in the Presence of Measurement Error: Case Study from the National Database of Nursing Quality Indicators® (NDNQI®)

Byron J. Gajewski^{1,3}, Robert Lee², and Nancy Dunton³

¹Department of Biostatistics, University of Kansas School of Medicine, Kansas City, KS, USA, 66160

²Department of Health Policy and Management, University of Kansas School of Medicine, Kansas City, KS, USA 66160

³University of Kansas School of Nursing, Kansas City, KS, USA 66160

Abstract

Data Envelopment Analysis (DEA) is the most commonly used approach for evaluating healthcare efficiency (Hollingsworth, 2008), but a long-standing concern is that DEA assumes that data are measured without error. This is quite unlikely, and DEA and other efficiency analysis techniques may yield biased efficiency estimates if it is not realized (Gajewski, Lee, Bott, Piamjariyakul and Taunton, 2009; Ruggiero, 2004). We propose to address measurement error systematically using a Bayesian method (Bayesian DEA). We will apply Bayesian DEA to data from the National Database of Nursing Quality Indicators® (NDNQI®) to estimate nursing units' efficiency. Several external reliability studies inform the posterior distribution of the measurement error on the DEA variables. We will discuss the case of generalizing the approach to situations where an external reliability study is not feasible.

Keywords

Healthcare Efficiency; Simulation; DEA; Bayesian; Quality Improvement

1. INTRODUCTION

Data Envelopment Analysis (DEA) is the most commonly used approach for evaluating healthcare efficiency (Hollingsworth, 2008). Further, DEA has been used to calculate a composite score for multivariate quality indicators (QIs) (Shwartz, Burgess, and Berlowitz, 2009).

DEA is attractive for several reasons. It does not require a complete specification of a functional form; it does not require the specification of an error distribution; and it easily handles cases with multiple inputs and multiple outputs.

However, a long-standing concern is that DEA assumes that data are measured without error. This is quite unlikely, and DEA and other efficiency analysis techniques may yield biased efficiency estimates in the presence of measurement error (Gajewski, Lee, Bott, Piamjariyakul and Taunton, 2009; Ruggiero, 2004). We propose to address the issue of

⁺ Address for correspondence: Department of Biostatistics, University of Kansas School of Medicine, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS, USA 66160, Phone: 913-588-1603, Fax: 913-588-0252, bgajewski@kumc.edu.

measurement error systematically using a Bayesian method (Bayesian DEA). An important advantage of Bayesian DEA is that it provides a general framework for a “stochastic” view of efficiency studies that can be used with other analytic approaches. Bayesian DEA extends an algorithm of Gajewski et al (2009) in that it accounts for the epistemic uncertainty in measurement error.

We will apply Bayesian DEA to data from the National Database of Nursing Quality Indicators® (NDNQI®) to estimate nursing units’ efficiency and composite QI (multiple input-output DEA). The posterior distribution of the measurement error on all of the variables used in the DEA is informed by data from several external reliability studies. The variables we will consider include staffing levels (for various skills), patient days, patient fall rates, and hospital-acquired pressure ulcer (bedsores) rates. We apply Bayesian DEA and classical DEA to this data and use correlation and 95% credible intervals to compare their efficiency scores (composite QIs). We will estimate of the impact of ignoring measurement error on classical DEA.

We will discuss the case of generalizing the approach to situations where an external reliability study is not feasible. In this case we discuss the elicitation of prior distributions of measurement error from experts as a promising option. Bayesian DEA provides the healthcare industry (and its stakeholders) unbiased estimates of efficiency and an approach that correctly distinguishes Decision Making Units (DMUs).

1.1 MOTIVATION

The United States spends nearly twice as much per person on medical care as other developed nations, yet has mediocre health outcomes (Anderson, Bianca, & Frogner, 2008). Improving health care efficiency is vital for the economic well-being of all Americans (The White House, 2011). Yet to improve efficiency analysts must be able to reliably identify efficient providers or types of care. Unfortunately, the ability of analysts to do so is not clear, as their ability to distinguish differences in efficiency from measurement error is suspect (Burgess, 2006). Measurement error has the greatest potential to produce misleading results when analysts use DEA (the most widely used approach), but the problem affects all forms of efficiency analysis (Hollingsworth, 2008).

The impact of measurement error is well-known for some statistical methods. For example, consider simple linear regression in which a single dependent variable (y) is regressed on a single predictor variable (x). If y is measured with error it impacts the precision of the regression parameter estimates. This is not really a problem since we can just take more measures. But we still obtain unbiased estimates of the parameter estimates. A more serious issue is when x is measured with error because it causes biased regression parameter estimates. A solution to this problem is to perform an external reliability study for estimating the measurement error for x and then adjust this bias using simulation (e.g., Apanasovich, Carroll, and Maity, 2009).

1.2. OUR QUESTION

The important goal of measuring healthcare efficiency via DEA is complicated because there are multiple inputs and multiple outputs. So it is not clear how the precision and/or the bias are affected. Therefore, in this paper we wish to address the following question: “Can we **reliably** estimate healthcare quality and more specifically efficiency?” We will address this question with data from the National Database of Nursing Quality Indicators (NDNQI). Our approach extends the methodology that focuses solely on regression (e.g. Buonaccorsi, 2010).

1.3. OUTLINE

In Section 2 we provide background on hospital report cards with a focus on the NDNQI database. We indicate that not only is DEA useful for estimating healthcare efficiency, it can be used for benchmarking. In Section 3 we provide a review of several studies of nursing efficiency. This provides a basis for the rational of our choice of variables for our efficiency case study using the NDNQI data. We discuss the measurement error issues in the NDNQI and report the reliability of our variables used as inputs and outputs into the DEA. In Section 4 we describe classical DEA algorithm and supply an extension that adjusts for measurement error using Bayesian principles. In Section 5 we provide details of the results of the case study. In general, an external reliability study may not be feasible, so we discuss what to do in this case in Section 6. We provide concluding remarks in Section 7.

2. BACKGROUND ON HOSPITAL REPORT CARDS & NDNQI DATASET

Since 1998, the National Database of Nursing Quality Indicators (NDNQI) has been providing hospitals report cards that are indicators of nursing unit's quality of care. Starting with 23 hospitals in 1999, it has grown to over 1,800 hospitals housing over 17,000 nursing units. Data are entered via a secure website and the NDNQI gives back benchmarking reports to each of the hospitals for them to understand how they are doing on nursing quality indicators relative to their peers. Hospitals mainly use the information from these reports for quality improvement purposes, for example, implementing new fall risk assessment procedures or preventing pressure ulcers.

The types of reports considered at NDNQI include univariate reports (Gajewski, Mahnken, & Dunton, 2008) as well as multivariate (i.e. composite indicators) (Gajewski, & Dunton, in press). Rather innovatively, Shwartz, Burgess, & Berlowitz (2009) consider using DEA for healthcare report cards. They argue that a DEA approach is a benefit-of-the-doubt approach for calculating a composite measure of quality. Therefore it is our contention that DEA be used for both measuring healthcare efficiency and healthcare quality report cards. However, before applying DEA to NDNQI, we need to choose input and output variables that are appropriate for application to nursing data. A brief review of the literature will elucidate our model development.

3. REVIEW OF NURSING UNITS' EFFICIENCY STUDIES AND CHOICE OF VARIABLES

Data and measurement issues for efficiency analysis require careful preparation. Coelli et al, 2005 mention two important steps, establishing the (1) micro-DEA approach and (2) macro-DEA approach. The micro-DEA approach refers to actually choosing the algorithm after the variables are chosen. For example, constant returns to scale or variable returns are common options. We will choose the classic Charnes-Cooper-Rhodes (CCR) approach (Cooper et al, 2004). The more important decision is the macro-DEA which involves choosing the actual input and output variables that go into the DEA.

Coelli et al (2005) defines five general categories of inputs: (1) labor is the most important (e.g. staff hours); (2) capital (e.g. beds, computers, technology); (3) energy; (4) materials; and (5) purchased services. Energy, materials, and purchased services are usually collapsed into an "other" category. In our case we will see that labor, such as nursing staff, is our key input variable. In the healthcare industry it is much more complex to correctly define outputs (Halsteinli, Kittelsen, Magnussen, 2010). To understand the difficulty, consider an industry that has tangible goods (e.g. number of computers produced). It is very easy to count the output here. In healthcare we can count the number of patient days, but need to attach some measure of the quality of care.

In Table 1 we list several nursing-focused hospital efficiency studies from 2006-2010. We have tabulated each of the studies in terms of unit of analysis, input variables, and output variables. All of the studies were at the hospital level of measurement except one which was at the nursing unit level (although had only 30 units). All of the studies included nursing labor with one broken down into Registered Nurse hours per patient day (X_1), Licensed Practical Nurse hours per patient day (X_2), and Unlicensed personnel hours per patient day (X_3). Some of the studies also included physicians and dentists. Capital examples included counts of beds and equipment specific to the problem (e.g. dialysis machines). One study had operating expenses, usually a hidden piece of information in hospital data. The outputs varied from values that were not quality specific (e.g. number of tests) to outputs that are measures of *nursing quality* specific (fall rates, satisfaction, etc) and case mix index adjusted.

For our study, most similar to Mark et al. (2009), we define inputs and outputs from NDNQI data. First, for inputs we have solely nursing labor: registered nurse hours per patient day (RNHPPD), licensed practicing nurse hours per patient day (LPNHPPD), and unlicensed assistive personnel hours per patient day (UAPHPPD). We have very little capital information at NDNQI, for example we do not have number of beds available to us. For outputs we focus on nursing related quality: fall rates and pressure ulcer rates. We will appropriately invert them so that they are on a positive scale, necessary for DEA. We address risk adjustment by analyzing unit level data (e.g. critical care, medical, surgical, combined medical and surgical, and step-down). We will focus on the 1,549 combined medical-surgical units that we have complete annual data from latter two quarters of 2009 and first two quarters of 2010 (fiscal year). In Table 2 we summarize the data that goes into these inputs and outputs.

4. DEVELOPING THE BAYESIAN APPROACH WITH CASE STUDY

The aims of this paper are to (1) develop a statistical approach to DEA that incorporates measurement error and (2) apply the Bayesian DEA approach and compare its performance to fixed (classical) DEA. Therefore we need information about measurement error. Fortunately we have it since NDNQI systematically performs reliability studies. Specifically we desire reliability studies for the DEA inputs (X 's) and the DEA outputs (Y 's). For the X 's we have reliability studies for the staffing variables (Klaus et al, in press) and for the patient days (Simon et al, 2011). For the Y 's we have reliability studies for the falls (Simon et al, in review) and pressure ulcer rates (Hart et al, 2006; Hart et al, 2010; Bergquist-Beringer et al, 2011; and Gajewski et al, 2007).

4.1 Mathematical Details of DEA

In the case in which inputs and outputs are measured without error, it is a straightforward linear programming problem to estimate the efficiency scores (Cooper et al, 2007, p. 23). Using Cooper's language, for each nursing unit (in general, DMU), Let $i=1,2,3,\dots,n$. Find the input "weights" (v_k) ($k=1,\dots,m$) and the output "weights" (u_l) ($l=1,\dots,s$) such that

$$\text{maximizes } \theta_i = (\sum u_l Y_{li}) / (\sum v_k X_{ki}) \quad (1)$$

with the constraint that $(\sum u_l Y_{li}) / (\sum v_k X_{ki}) \leq 1$ (for all $i=1,2,3,\dots,n$), v 's ≥ 0 , and u 's ≥ 0 .

The constraints imply that the ratio of "virtual output" versus "virtual input" should not exceed 1 for every nursing unit. The objective is to obtain weights (v 's) and (u 's) that maximize the ratio of the nursing unit being evaluated. In the case of $\theta_i=1$ the nursing unit is "fully efficient," whereas $\theta_i=0$ identifies a nursing unit that is "fully inefficient."

Classically this framework of DEA is non-stochastic, therefore if there is no measurement error θ_j is perfectly measured. But in our case measurement error in the inputs and outputs will alter θ_j . For the sake of nomenclature we will call the data measured without error Y^* and X^* for the respective outputs and inputs. We will use reliability studies to obtain posterior distributions of $Y^*|Y, \phi_1$ and $X^*|X, \phi_2$ where X and Y are the observed outputs and inputs and ϕ_1 and ϕ_2 are parameters (technically they are point estimates from these studies) learned from the external reliability studies. Using the posterior distributions of Y^* and X^* we can gain a posterior distribution of the efficiency scores which we will call $\theta^*_{j|Y, X, \phi_1, \phi_2}$, where the '*' helps clearly indicate that data adjusted for measurement error are used to calculate this posterior distribution. This will be very useful for calculating how far off we are in efficiency scores if we ignore the measurement error. For example we can get the Bias = $\theta_j - E(\theta^*_{j|Y, X, \phi_1, \phi_2})$ and Variance = $\text{Var}(\theta^*_{j|Y, X, \phi_1, \phi_2})$, which calculations will be demonstrated via case study on NDNQI data. The mean squared error is $\text{MSE} = \text{Bias}^2 + \text{Variance}$.

4.2 Efficiency Scores for Observed Data

For the $n=1,549$ medical-surgical units described in Table 2, we calculated each of the nursing unit's efficiency scores which are displayed in Figure 1. The distribution is slightly skewed to the left and the percentiles of efficiency indicates that 95% of the efficiency scores are between 0.4255–0.9145, the middle 50% scores are between 0.5604–0.7165, and 50% of the efficiency scores are below 0.6351. Thus a majority of the nursing units of medical-surgical type are above 50% efficient.

4.3 What about Measurement Error?

How reliable are these percentiles? This is addressed through the adjustment of measurement error. Some notation is needed to setup the algorithm. Let $\mathbf{Z}_j$ be a vector of data for the j^{th} nursing unit. Let the input and the outputs for the DEA be $\mathbf{X}_j = \mathbf{f}_1(\mathbf{Z}_j)$ and $\mathbf{Y}_j = \mathbf{f}_2(\mathbf{Z}_j)$. Essentially what we have is reliability studies on $\mathbf{Z}_j$, which after transformation gives us the reliability of the vectors $\mathbf{X}_j$ and $\mathbf{Y}_j$ that go into the DEA. In our medical-surgical units example we have individual elements as $Z_1 = \text{RN hours}$, $Z_2 = \text{LPN hours}$, $Z_3 = \text{UAP hours}$, $Z_4 = \text{patient days}$, $Z_5 = \text{pressure ulcers}$, $Z_6 = \text{falls}$, and $n = \text{number of patients assessed}$. Thus the transformation to inputs and outputs is $X_1 = Z_1/Z_4$, $X_2 = Z_2/Z_4$, $X_3 = Z_3/Z_4$, $Y_1 = \max - Z_5/Z_4 * 1000$, and $Y_2 = 1 - Z_6/n$, and max is the maximum fall rate ($Z_6/Z_4 * 1000$) across all units.

4.3.1 Lognormal model (Inputs: RNHPPD, LPNHPPD, UAPHPPD)—Several of the variables in the vector $\mathbf{Z}$ have measurement error models on the lognormal scale. Specifically we use the notation that Z_j is the observed value, $Z^*_{j|}$ is the unobserved true value, and ϵ_j is the multiplicative error. The lognormal measurement error model is written

$$\log(Z_j) = \log(Z^*_{j|}) + \log(\epsilon_j), \quad (2)$$

where we assume that $\log(Z^*_{j|}) \sim N(0, \sigma^2_{Bj})$ and $\log(\epsilon_j) \sim N(0, \sigma^2_{Wj})$; σ^2_{Bj} represents the between variance of the log of the latent variable and σ^2_{Wj} is the variance of the measurement error of the log of the observed variable. The intraclass correlation coefficient (ICC) under this parameterization is $\text{ICC} = \sigma^2_{Bj} / (\sigma^2_{Bj} + \sigma^2_{Wj}) = \sigma^2_{Bj} / \sigma^2_{Tj}$. Using lognormal ratios it is easily shown that the posterior distribution

$$Z^*_{j|} | Z_j, \sigma^2_{Wj} \sim \text{LN}(\log(Z_j), \sigma^2_{Wj}), \quad (3)$$

where "LN" is the lognormal distribution. To obtain this distribution the "sufficient information" for obtaining the posterior distribution in Equation (3) is Z_j , ICC_j , and σ^2_{Tj} .

(Note that $\sigma^2_{w_j} = (1 - \text{ICC}_j) \sigma^2_{T_j}$). Using Equation (3) then we can also calculate the reliability of the ratio of two variables measured with error, for example, suppose we want the true distribution of RNHPPD which is the ratio of the variables for $j=1$ and $j=4$, then we get

$$Z^*_1/Z^*_4|Z_1, Z_4, \sigma^2_{w1}, \sigma^2_{w4} \sim \text{LN}(\log(Z_1) - \log(Z_4), \sigma^2_{w1} + \sigma^2_{w4}). \quad (4)$$

Table 3 shows what we call “sufficient information” for estimating the measurement error for the four Z variables specifically we need total and within variances. As an example, consider a case in which we have a unit that we observe $Z_4=5,067$ patient days and $Z_1=27,958$ RN hours, which results in $X_1=Z_1/Z_4=5.52$ RN hours per patient day. Using the results in Equations (3) and (4) as well as the “sufficient parameters” in Table 3, we gain the posterior distributions shown in Figure 2.

Rather than the closed form distributions, one can also use simulation to generate approximations to the posterior distribution. This will be useful later, so we illustrate how to do so on the lognormal examples. Suppose we have B simulations, then one can simulate the b^{th} iteration of Equation (3) by simulating

- $\log(\varepsilon^{(b)}_j) \sim N(0, \sigma^2_{w_j})$, and then (5)
- $Z^{*(b)}_j = Z_j \varepsilon^{(b)}_j$

Repeat this expression in Equation (5), say $B=1000$ times and obtain the Kernel estimated smoothed values in Figure 3, closely resembling the results in Figure 2.

4.3.2 Binomial model 1 (Output: Pressure Ulcer Rate)—The basic model for understanding the measurement error for the pressure ulcer rates begins at the individual pressure ulcer and establishing sensitivity and specificity from a survey of pressure ulcer pictures (Hart et al, 2006; Hart et al, 2010). Let y_i be the observed pressure ulcer status for individual i . Let y^*_i be the true pressure ulcer status for individual i . Then using the reliability study’s point estimates, we estimate the specificity to be $P(y^*_i=0|y_i=0)=0.98$ and the sensitivity to be $P(y^*_i=1|y_i=1)=0.81$. Now convert the individual model to a unit model. First, suppose we observe Z_5 pressure ulcers and n patients assessed ($n-Z_5$ patients have no pressure ulcers). Then based on the above model we can see $U^*_5 \sim \text{Binomial}(0.98, n-Z_5)$, the number of false negatives, and $V^*_5 \sim \text{Binomial}(0.81, Z_5)$, the number of true positives. Combining this information

$$Z^*_5 = (n - Z_5 - U^*_5) + V^*_5. \quad (6)$$

Using a numerical example, suppose we observe $n=113$ patients with $Z_5=4$ having pressure ulcers (PrU Rate= 3.54%). Suppose we have B simulations, then one can simulate the b^{th} iteration by simulating $U^{*(b)}_5 \sim \text{Binomial}(0.98, 113-4)$, $V^{*(b)}_5 \sim \text{Binomial}(0.81, 4)$, & $Z^{*(b)}_5 = (113-4 - U^{*(b)}_5) + V^{*(b)}_5$. Repeat this expression in Equation (6), say $B=1000$ times to obtain the histogram in Figure 4.

4.3.3 Binomial model 2 (Output: Fall Rate)—The reliability of falls can be summarized by the famous quote “If a tree falls in a forest and no one is around to hear it, does it make a sound?” (see for example Wikipedia, 2011). Unlike pressure ulcer rates, where we know how many patients we did not observe a pressure ulcer, we don’t know the number that was not observed to fall. Therefore our fall rate measurement error model involves sensitivity only (Simon, et al, in review). For an individual who was observed to

fall, let f_i be the observed fall status and f_i^* be the true fall status for individual i . Then the sensitivity is $P(f_i^*=1|f_i=1)=0.962$. The model becomes as follows. Suppose we observe Z_6 falls, then the number of true positives is $Z_6^* \sim \text{Binomial}(0.962, Z_6)$. Therefore, suppose we have B simulations, then one can simulate the b^{th} iteration by simulating

$$\bullet \quad Z_6^{*(b)} \sim \text{Binomial}(0.962, Z_6), \text{ \& \quad (7)}$$

- Log-normal model patient days (via Equation 5).

For example, suppose we observe 10,616 patient days and $Z_5=36$ falls (Fall Rate= 3.39). Using Equation 7, simulate true falls and patient days and then report their posterior distributions as shown in Figure 5.

4.3.4 A Bayesian Approach for Estimating DEA Accounting for Measurement Error—By combining the steps above, we obtain an overall methodology that obtains the posterior distribution of $\theta_i^* | Y, X, \varphi_1, \varphi_2$.

- Step 0: Set $b=0$
- Step 1: Set $b=b+1$
- Step 2: Simulate $Z^{*(b)}$ using the approaches previously outlined (Equations 5, 6, & 7).
- Step 3: Calculate inputs ($X^{*(b)}$) and outputs ($Y^{*(b)}$) using $Z^{*(b)}$ from Step 2.
- Step 4: Run DEA algorithm on simulated data (Equation 1).
- Step 5: Save the efficiency scores for each unit ($\theta_i^{*(b)} | Y, X, \varphi_1, \varphi_2$), for $i=1, 2, 3, \dots, n$.
- Step 6: Go to Step 1 (stop when $b=1000$).
- Step 7: Summarize

In the final step we summarize the p^{th} percentile with a kernel density estimator, and call that $\theta^{*(p)} | Y, X, \varphi_1, \varphi_2$. Note that the p^{th} value is defined from the distribution of efficiency scores on the observed data. We can also estimate the difference in percentiles, for example, the difference in the 75%-tile and the 25%-tile as $\theta^{*(0.75)} | Y, X, \varphi_1, \varphi_2 - \theta^{*(0.25)} | Y, X, \varphi_1, \varphi_2$. To measure global efficiency we have the posterior distribution estimated as

$$\bar{\theta}_\bullet^{*(b)} | X, Y, \varphi_1, \varphi_2 = \sum_{i=1}^n \{ \theta_i^{*(b)} | X, Y, \varphi_1, \varphi_2 \} / n.$$

5. RESULTS OF BAYESIAN APPROACH VIA CASE STUDY

Figure 6 shows the posterior distribution of the efficiency scores for the 2.5%-tile, 25%-tile, 50%-tile, 75%-tile, and 97.5%-tile and the observed efficiency scores. It highlights the uncertainty in each of these percentiles. The half width of each of these 95% intervals gets larger as the percentile gets larger. The half widths for these percentiles are 0.10, 0.13, 0.15, 0.17, and 0.19 respectively. The biases for these cases are 0.05, 0.05, 0.08, 0.09, and 0.13. These correspond to a mean squared error of 0.0048, 0.0072, 0.0130, 0.0164, and 0.0257. Figure 7 summarizes the posterior distributions for the differences between 75%-tile and 25%-tile as well as differences between 97.5%-tile and 2.5%-tile. The posterior probability that the 25%-tile is bigger than 75%-tile is $\Pr(\{ \theta^{*(0.75)} | Y, X, \varphi_1, \varphi_2 - \theta^{*(0.25)} | Y, X, \varphi_1, \varphi_2 \} < 0) = 0.1370$, using a 0.05 cut-point, so they are not very different from each other. Conversely, $\Pr(\{ \theta^{*(0.975)} | Y, X, \varphi_1, \varphi_2 - \theta^{*(0.025)} | Y, X, \varphi_1, \varphi_2 \} < 0) < 0.001$, therefore in the percentiles further out we have detected differences at the 0.05 cut-point.

For overall average efficiency, the half-width is 0.04, the bias is 0.07, and the mean squared error is 0.0049.

6. GENERALIZING THE APPROACH TO SITUATIONS WHERE AN EXTERNAL RELIABILITY STUDY IS NOT FEASIBLE

The data from the NDNQI case study allows us to measure error of all of the variables via external reliability studies. However, in general, reliability studies may not be available and too costly to perform. In these cases we are in need of estimates of measurement error. In this case we will train experts in the definition of measurement error and build elicitation tools that will inform expert-based estimates of measurement error. For example, recruit six content experts to be interviewed regarding their opinion of the measurement error. Before the interview we would ask them to review relevant literature so that they are familiar with the variables. Then we would continue the interview, for example, we would ask them *why the error varies in their population or the targeted item under study*. To gauge our ability to train we would evaluate the experts' measurement error on a variable that we might know its measurement error. The iterative process involves: eliciting summaries of experts' measurement error, fitting a model to these summaries, and assessing the accuracy, and if inadequate repeat the process (O'Hagan et al, 2006). This summary for all variables (especially the ones with no external reliability studies) will be used for identifying measurement error to be programmed into Bayesian DEA.

7. DISCUSSION & CONCLUSION

We have demonstrated that we can reliably measure global efficiency. Our proposed methodology for adjusting for measurement error in DEA is broadly applicable. It can incorporate a number of distributions that are non-normal (e.g. lognormal and Binomial). The Bayesian approach via simulation allows for this flexibility. In this paper we compare Bayesian DEA to classical DEA, but because the Bayesian approach is used to modify the underlying data, the data can be analyzed in a number of ways. Once the modification is complete, the data can be analyzed using stochastic frontier analysis (Kumbhakar & Lovell, 2003), corrected ordinary least squares, or quantile regression. Our approach, therefore, applies to all of the widely-used strategies for assessing efficiency.

Our case study focuses on hospitals, but the tools and techniques that we propose here can be fully applicable to other healthcare settings. For example, DEA has been used to study efficiency in nursing homes, health maintenance organizations, and a variety of other settings (Lenard and Shimshak, 2009). In fact, our method is applicable for all activities analyzed with DEA. If measurement error is substantially greater in healthcare than in other activities evaluated with DEA, then the DEA scores estimated with classical deterministic models would be even less reliable.

Our case study indicates that even with 'substantially reliable' data, DEA can vary in precision: (1) differences in 2.5%-tile and 97.5%-tile are pronounced; (2) but differences in 75%-tile and 25%-tile not as pronounced. However, this quantification of the measurement error at the efficiency stage guides us in saying what units are truly different. In our view this is a significant discovery. For our data set, we cannot statistically distinguish the DEA scores at the 25th and the 75th percentiles. This finding has big implications about the validity of comparisons of the DEA scores of individual DMUs. For example, for any pair of DMU's DEA score between the 25th and 75th percentiles, we cannot make any worthwhile evaluation about their comparative efficiencies. Further, we have quantified that the overall estimate of efficiency is biased by .06. Therefore we can reliably measure efficiency in medical-surgical units up to a point. A current limitation of our approach is that it would be

helpful to incorporate full uncertainties of external reliability studies; currently we incorporate only the point estimate.

Our research has motivated two new questions: (1) can we longitudinally estimate healthcare efficiency (e.g. stock market)? & (2) can we integrate this approach with the approach of Gajewski et al (2009) and Barnum et al (2012)? Further research will address both questions.

Clearly DEA is a useful tool for increasing the efficiency of healthcare. Our contention is that it needs to incorporate measurement error. We addressed this by developing a Bayesian DEA approach that adjusts for measurement error. An adjusted DEA allows researchers to have tools for properly assessing the efficiency of hospitals, nursing homes, and many healthcare settings. Regardless of the political climate of healthcare in the next several years, researchers and policy makers will want the correct picture regarding the healthcare efficiency in the US, which is provided by our methodology.

Acknowledgments

This work was supported in part by the National Database of Nursing Quality Indicators® (NDNQI®), from a contract with the American Nurses Association® (first and third authors) and Frontiers: The Heartland Institute for Clinical and Translational Research CTSA UL1RR033179 which is now UL1TR000001 (first author). The contents are solely the responsible of the authors and do not necessarily represent the official views of the NIH and ANA.

REFERENCES

- Anderson GF, Bianca K, Frogner BK. Health spending in OECD countries: obtaining value per dollar. *Health Affairs*. 2008; 27(6):1718–1727. [PubMed: 18997231]
- Apanasovich TV, Carroll RJ, Maity A. SIMEX and variance estimation in semi-parametric measurement error models. *Electronic Journal of Statistics*. 2009; 3:318–348. [PubMed: 19609371]
- Barnum D, Gleason J, Karlaftis M, Schumock G, Shields K, Tandon S, Surrey Walton S. Estimating DEA confidence intervals with statistical panel data analysis. *Journal of Applied Statistics*. 2012; 39(4):815–828.
- Buonaccorsi, JP. *Measurement Error: Models, Methods, and Applications*. New York, NY: CRC Press; 2010.
- Bergquist-Beringer S, Gajewski B, Dunton N, Klaus SK. The reliability of the National Database of Nursing Quality Indicators (NDNQI) pressure ulcer indicator: A triangulation approach. *Journal of Nursing Care Quality*. 2011; 26(4):292–301. [PubMed: 21407089]
- Burgess, JF. *Productivity analysis in health care*. In: Jones, A., editor. *The Elgar Companion to Health Economics*. Cheltenham: Edward Elgar; 2006.
- Coelli, TJ.; Rao, DSP.; O'Donnell, CJ.; Battese, GE. *An Introduction to Efficiency and Productivity Analysis*. 2nd Edition. New York, NY: Springer; 2005.
- Cooper, W.; Seiford, L.; Tone, K. *Data envelopment analysis: a comprehensive text with modes, applications, references and DEA-solver software*. London: Kluwer Academic Publishers; 2007.
- Gajewski BJ, Dunton N. Identifying individual changes in performance with composite quality indicators while accounting for regression-to-the-mean. *Medical Decision Making*. (in press).
- Gajewski BJ, Hart S, Bergquist S, Dunton N. Inter-rater reliability of pressure ulcer staging: ordinal Probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine*. 2007; 26(25):4602–4618. [PubMed: 17393413]
- Gajewski BJ, Lee R, Bott M, Piamjariyakul U, Taunton RL. On estimating the distribution of data envelopment analysis efficiency scores: an application to nursing homes' care planning process. *Journal of Applied Statistics*. 2009; 36(9):933–944.
- Gajewski BJ, Mahnken JD, Dunton N. Improving quality indicator report cards through Bayesian modeling. *BMC Medical Research Methodology*. 2008; 8:77. <http://www.biomedcentral.com/1471-2288/8/77>. [PubMed: 19017399]

- Groff JE, Lien D, Su J. Measuring efficiency gains from hospital mergers. *Research in Healthcare Financial Management*. 2007; 11(1):77–90.
- Halsteinli V, Kittelsen SA, Magnussen J. Productivity growth in outpatient child and adolescent mental health services: the impact of case-mix adjustment. *Social Science & Medicine*. 2010; 70(3):439–446. [PubMed: 19926189]
- Hart S, Bergquist S, Gajewski B, Dunton N. Reliability testing of the National Database of Nursing Quality Indicator's pressure ulcer indicator. *Journal of Nursing Care Quality*. 2006; 21(3):256–265. [PubMed: 16816607]
- Hart S, Bergquist S, Gajewski B, Dunton N. Reliability testing of the National Database of Nursing Quality Indicator's pressure ulcer indicator. *Journal of Nursing Administration*. 2010; 40(10):S16–S25. [PubMed: 20859098]
- Hollingsworth B. The measurement of efficiency and productivity of health care delivery. *Health Economics*. 2008; 17(10):1107–1128. [PubMed: 18702091]
- Klaus S, Gajewski B, Potter C, Dunton N. Reliability of the nursing care hour measure. *International Journal of Nursing Studies*. (in press).
- Kontodimopoulos N, Niakas D. A 12-year Analysis of Malmquist Total Factor Productivity in dialysis facilities. *Journal of Medical Systems*. 2006; 30:333–342. [PubMed: 17068996]
- Kontodimopoulos N, Moschovakis G, Aletras VH, Niakas D. The effect of environmental factors on technical and scale efficiency of primary health care providers in Greece. *Cost Effectiveness and Resource Allocation*. 2007; 5(1):14–24. [PubMed: 18021419]
- Kumbhakar, SC.; Lovell, CAK. *Stochastic Frontier Analysis*. Cambridge: Cambridge Press; 2003.
- Lenard LM, Shimshak DG. Benchmarking nursing home performance at the state level. *Health Services Management Research*. 2009; 22:51–61. [PubMed: 19401498]
- Mark BA, Jones CB, Lindley L, Ozcan YA. An examination of technical efficiency, quality, and patient safety in acute care nursing units. *Policy Polit Nurs Pract*. 2009; 10(3):180–186. [PubMed: 20008398]
- O'Hagan, A.; Buck, C.; Daneshkhah, A.; Eiser, J.; Garthwaite, P.; Jenkinson, D.; Oakley, J.; Rakow, T. *Uncertain judgments: Eliciting experts' probabilities*. New York: Wiley; 2006.
- Puenpatom RA, Rosenman R. Efficiency of Thai provincial public hospitals during the introduction of universal health coverage using capitation. *Health Care Management Science*. 2008; 11:319–338. [PubMed: 18998592]
- Ruggiero J. Data envelopment analysis with stochastic data. *Journal of the Operational Research Society*. 2004; 55:1008–1012.
- Shwartz M, Burgess JF, Berlowitz. Benefit-of-the-doubt approaches for calculating a composite measure of quality. *Health Serv Outcomes Res Method*. 2009; 9:234–251.
- Simon M, Yankovskyy E, Klaus S, Gajewski B, Dunton N. Midnight census revisited: reliability of patient day measurement and its impact on nursing quality indicators. *International Journal of Nursing Studies*. 2011; 48(1):56–61. [PubMed: 20673896]
- Simon M, Klaus S, Gajewski B, Dunton N. Falls reliability study. (in review).
- The White House, Health Care. [accessed on October 18, 2011] <http://www.healthcare.gov/law/about/index.html>
- Valdmanis VG, Rosko MD, Mutter RL. Hospital quality, efficiency, and input slack differentials. *Health Services Research*. 2008; 43(5):1830–1848. [PubMed: 18783457]
- Wikipedia. [accessed November 22, 2011] 2011. http://en.wikipedia.org/wiki/If_a_tree_falls_in_a_forest

Figure 1.

The DEA scores using the classical algorithm described in Equation (1) for $n=1,549$ medical-surgical units. Key summary statistics: 2.5%-tile=0.4255, 25%-tile=0.5604, 50%-tile=0.6351, 75%-tile=0.7165, and 97.5%-tile=0.9145.

Figure 2.

Consider a case in which we have a unit that we observe $Z_4=5,067$ patient days and $Z_I=27,958$ RN hours, which results in $X_I=Z_I/Z_4=5.52$ RN hours per patient day. Using the results in Equations (3) and (4) as well as the “sufficient parameters” in Table 3, we gain the posterior distributions shown here.

Figure 3. Repeat this expression in Equation (5), $B=1,000$ times and obtain the Kernel estimated smoothed values below, closely resembling the results in Figure 2.

Figure 4.

Suppose we observe $n=113$ patients with $Z_5=4$ having pressure ulcers (PrU Rate= 3.54%). Suppose we have B simulations, then one can simulate the b^{th} iteration by simulating $U^{*(b)}_5 \sim \text{Binomial}(0.98, 113-4)$, $V^{*(b)}_5 \sim \text{Binomial}(0.81, 4)$, & $Z^{*(b)}_5 = (113-4 - U^{*(b)}_5) + V^{*(b)}_5$. Repeat this expression in Equation (6), $B=1000$ times to obtain this histogram.

Figure 5. Suppose we observe 10,616 patient days and $Z_5=36$ falls (Fall Rate= 3.39). Using Equation 7, simulate true falls and patient days and then report their posterior distributions as shown here.

Figure 6. Posterior distributions of efficiency scores.

Figure 7.
Posterior distributions of efficiency scores for differences.

Figure 8.
Posterior distributions for overall efficiency.

Table 1

Several healthcare efficiency studies from 2005–2010 involving nursing.

Study	Unit of Analysis	Inputs	Outputs
Mark et al. (2009)	Nursing Unit	Labor: RNHPPD, LPNHPPD, UAPHPPD Capital: #Beds Other: Operating expenses	Adjusted discharges, Patient Satisfaction, Medication rate, Fall rate
Valdmanis et al (2008)	Hospitals	Labor: FTE RNs, FTE LPNs, FTE other Capital: Bassinets, acute beds, other beds Other: None	total outpatient visits, total surgeries, total births, Medicare Case Mix Index adjusted admissions, total other patient days
Kontodimopoulos et al (2007)	Hospitals	Labor: Physicians, nursing/paramedical, administrative/support staff Capital: None Other: None	#Diagnostic Tests
Puenpatom & Rosenman (2008)	Hospitals	Labor: Physicians, nursing/paramedical, administrative/support staff Capital: None Other: None	#Diagnostic Tests
Groff et al (2007)	Hospitals	Labor: FTE Physicians, dentists, FTE RN, FTE LPN, FTE other personnel Capital: # beds, # high technology, services Other: None	Inpatient days, outpatient visits, FTE trainees, Standby services
Kontodimopoulos et al (2006)	Hospitals	Labor: Nurses Capital: #Dialysis Machines Other: None	# of discharges, medical research

Table 2

Summary of raw variables, indicators, and for NDNQI efficiency study. There are 1,549 units and all are combined medical and surgical. The parenthetical (I) and (O) indicate the form of the variables in input and output form.

Variable	Label	Min	Max	Mean	Sd
RN Hours	Z_1	10838	221275	46053	17941
LPN Hours	Z_2	0	45755	3399	5380
UAP Hours	Z_3	0	87856	20463	10360
Patient Days	Z_4	1312	24196	8195	2936
#Falls	Z_5	0	101	29.8	16.2
#Hospital Pr Ulcers	Z_6	0	28	2.7	3.1
Pts Assessed for PrU	n	9	535	91.3	44.3
RNHPPD (I)	$X_1=Z_1/Z_4$	2.43	14.33	5.75	1.39
LPNHPPD (I)	$X_2=Z_2/Z_4$	0.00	4.52	0.43	0.65
UAPHPPD (I)	$X_3=Z_3/Z_4$	0.00	7.06	2.51	0.90
Fall Rate	Z_6/Z_4*1000	0.00	13.34	3.63	1.59
PrU Rate	Z_6/n	0.00	0.34	0.03	0.03
max-Fall Rate (O)	$Y_1=\max-Z_6/Z_4*1000$	0.00	13.34	9.71	1.59
1-PrU (O)	$Y_2=1-Z_6/n$	0.66	1.00	0.97	0.03

\$watermark-text

\$watermark-text

\$watermark-text

Table 3
With Z_j , this is the “sufficient information” for estimating the posterior distribution of Z_j^* and ratios of any.

Variable	j	ICC _{j}	Study	σ^2_{Tj}	σ^2_{wj}
log(RN Hours)	1	0.95	Klaus et al, in press	0.15	0.01
log(LPN Hours)	2	0.93	Klaus et al, in press	16.40	1.15
log(UAP Hours)	3	0.94	Klaus et al, in press	0.44	0.03
log(Patient Days)	4	0.94	Simon et al, 2011	0.15	0.01