Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP06/2020

Aggregate Efficiency of Industry and its Groups: The case of Queensland Public Hospitals

Bao Hoang Nguyen and Valentin Zelenyuk

Date: April 2020

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Aggregate Efficiency of Industry and its Groups: The case of Queensland Public Hospitals

Bao Hoang Nguyen* Valentin Zelenyuk†

April 21, 2020

Abstract

In this paper, we explore the efficiency of different groups of hospitals in Queensland, Australia, focusing on teaching and non-teaching hospitals, by adapting the most recent developments on statistical analysis of aggregate efficiency. We focus on the two approaches: the bootstrap approach proposed by Simar and Zelenyuk (2007) and the central limits theorems recently developed by Simar and Zelenyuk (2018, 2020). To adapt these developments, we extend the central limit theorems to the context where there are several sub-groups in the population. Using real data on Queensland public hospitals, we found that teaching hospitals are significantly less efficient than non-teaching hospitals when benchmarking is done with respect to the constant returns to scale frontier, but are significantly more efficient when benchmarking with respect to the variable returns to scale frontier.

Keywords: Hospitals, Aggregate Efficiency, Envelopment Estimators, Bootstrap, Central Limit Theorems.

JEL Codes: C24, C61, I11, I18.

*School of Economics, University of Queensland, Brisbane, Qld 4072, Australia

†School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Qld 4072, Australia

1 Introduction

The focus of empirical analysis in this paper is to compare the efficiency of teaching and non-teaching hospitals. Teaching hospitals are sometimes expected to be less efficient than non-teaching hospitals because they have to utilise more resources to fulfil the educational mission. The higher level of resource utilisation in teaching hospitals is not only because more overheads and equipment are needed for teaching and research activities, but it is also due to the increased use of ancillary services from residents in training, who tend to keep patients longer and may over-prescribe medical tests (Cameron, 1985; Rich et al., 1990). Moreover, the productivity of medical staff in teaching hospitals is expected to be negatively impacted by their teaching mission, since they have to share their time between providing patient care services and training residents (Jensen & Morrisey, 1986; Hao & Pegels, 1994).

On the other hand, teaching hospitals are widely reputed for providing high-quality care and are the places where many innovative medical techniques are first discovered and applied (Ayanian & Weissman, 2002; Shahian et al., 2012). In addition, the involvement in teaching and research activities might also help medical staff to develop professional skills and to accumulate human capital, which in turn helps them to be more effective in treating patients. As a result, the effectiveness of diagnosis and treatment in teaching hospitals is expected to be higher than in non-teaching hospitals. With these two opposite driving forces, the relative efficiency between teaching and non-teaching hospitals is not known in prior, and thus is an empirical question.

The empirical evidence about the relationship is rather mixed in the literature. For example, Lehner and Burgess (1995) and Grosskopf et al. (2001) found that teaching hospitals are less efficient than non-teaching hospitals, but Nayar et al. (2013) found the opposite evidence, while Burgess and Wilson (1998) did not find any evidence about the difference in efficiency between teaching and non-teaching hospitals although all these studies were undertaken in the U.S. context. This paper aims to shed more light on this important debate in the literature. To do so, we adapt the most recent developments on statistical analysis of aggregate efficiency and apply them to the data on public hospitals in Australia.

1.1 The Importance of Measuring Aggregate Efficiency

Besides estimating the efficiency of individual production units, researchers often need to obtain measures of efficiency at more aggregate levels, for example, the efficiency of some groups of interest within an industry or the efficiency of the industry as a whole.

Hereafter, we will call these measures with a common name: group efficiencies and will distinguish between entire group efficiencies and sub-group efficiencies where needed.

A natural measure of group efficiencies is the simple average of individual efficiency scores of production units within the group. It is a very simple approach, yet its drawback is that it ignores the size and hence the economic influence or weight of each individual in the aggregate. To make this point more intuitive, let us consider a hypothetical industry, which includes one big firm and ten small firms. As is the reality, where many industries are dominated by a few firms, suppose the big firm accounts for 90% of the industry revenue, whereas each of the ten small firms makes up only 1% of the revenue. Let us further assume that the big firm is very efficient, say, 100% efficient, meanwhile all the small firms are only 50% efficient. If one looks at the simple average, the industry is only 55% efficient. Obviously, the simple average does not take into account the fact that the industry is dominated by a very efficient firm - the big firm. One way to account for the economic weight of each individual is to use the aggregate efficiency measure proposed by Färe and Zelenyuk (2003), extended by Simar and Zelenyuk (2007). As compared to the simple average, their approach is useful because it uses meaningful weights derived from the economic optimization principle to aggregate individual efficiency scores, thus it takes the relative economic importance of individual production units in the group into account.

1.2 Statistical Inference of Aggregates of Envelopment Estimates

It is well-known in the literature that the conventional statistical inferences fail to apply to envelopment estimators of aggregate efficiency because these estimators inherit the statistical properties from individual efficiency estimators, which are biased and the bias is of a higher order than the variance when the dimension of the production model (the number of inputs and outputs) increases (Kneip et al., 2015). Prior to this important work, Simar and Zelenyuk (2007) attempted to correct the bias and perform statistical inferences for the aggregate efficiency by adapting the subsample bootstrapping technique proposed by Kneip et al. (2008) to the context of the group-wise heterogeneity.

Recently, Simar and Zelenyuk (2018) extended results from Kneip et al. (2015) by developing central limit theorems for aggregate efficiency. They utilized appropriate techniques to correct bias and control convergence rates for both bias and variance. The new central limit theorem has opened the path for more precise and theoretically-grounded statistical inference for aggregate efficiencies.

It is important to note that the results in Simar and Zelenyuk (2018) were developed

only for the case of entire group aggregate efficiency. To adapt these developments, one needs to carefully extend some of the results to the case when there are different sub-groups in the population, which appear to be new and thus is a modest theoretical novelty of this paper. Moreover, this paper is the first real use of these methods (with some novel extensions) for an empirical study.

1.3 Empirical Findings

Using the data on public hospitals in Queensland, Australia, we found that the conclusions about the relative efficiency between teaching and non-teaching hospitals dramatically depended on the adopted reference technology. Specifically, when benchmarking to the constant returns to scale (CRS) frontier, teaching hospitals are significantly less efficient than non-teaching hospitals. However, teaching hospitals are significantly more efficient than non-teaching hospitals when benchmarking to the variable returns to scale (VRS) frontier. The difference in these opposite conclusions is largely explained by the diseconomies of scales of teaching hospitals, which are primarily large (and perhaps "too large"?) hospitals. This raises a few natural questions for policy makers: Should the teaching hospitals be so large? Or, more generally, should the government, instead of expanding already large hospitals, invest into building new hospitals that are near the socially-optimal scale of operations? Indeed, the hospitals operating near the optimal scale must be more capable of expanding their operations (and at greater level of productivity) when facing sudden needs for its services, e.g., as during pandemics or other healthcare challenges to our society.

This paper is organized as follows. Section 2 discusses the methodologies. Section 3 describes the data and variables. Section 4 discusses the results, and Section 5 provides concluding remarks.

2 Methodology

2.1 Individual and Aggregate Efficiency

Let us consider a production technology, in which Decision Making Units (DMUs) use p inputs denoted as a p-dimensional column vector $x \in \Re^p_+$ to produce q outputs denoted as a q-dimensional column vector $y \in \Re^q_+$. The corresponding prices are denoted by row vectors $v \in \Re^p_{++}$ and $w \in \Re^q_{++}$, respectively. We assume that the technology can be

characterized by a technology set defined as follows

$$\Psi = \{(x, y) \in \Re_+^p \times \Re_+^q \mid x \text{ can produce } y\}, \tag{1}$$

or equivalently an output set defined as

$$P(x) = \{y \in \Re_+^q \mid x \text{ can produce } y\}, \tag{2}$$

and the technology or the production frontier is then defined based on the technology set

$$\Psi^{\partial} = \left\{(x, y) \in \Psi \;\middle|\; \left(\delta^{-1}x, \delta y\right) \notin \Psi, \forall \delta > 1\right\}. \tag{3}$$

We also assume that the standard regularity conditions of production theory (Shephard, 1953, 1970; Färe & Primont, 1995) are satisfied. Specifically,[1]

A1. "No Free Lunch", i.e., $y \notin P(\mathbf{0}_p),\ \forall y \geq \mathbf{0}_q$ and $y \neq \mathbf{0}_q$.

A2. "Producing Nothing is Possible", i.e., $\mathbf{0}_q \in P(x),\ \forall x \in \Re_+^p$.

A3. "Boundedness of the Output Sets", i.e., $P(x)$ is a bounded set for all $x \in \Re_+^p$.

A4. "'Closedness' of the Technology set", i.e., Ψ is a closed set.

A5. "Strong Disposability of All Inputs and Outputs", i.e, $(x_0, y_0) \in \Psi \Rightarrow (x, y) \in \Psi, \forall x \geq x_0, y \leq y_0$.

Farrell type efficiency measures appear to be the most widely-used measures of efficiency in the literature and the most general measure of this type, which encompasses other conventional Farrell-type measures as its multiplicative components, is Farrell profit efficiency (recently developed by Färe et al., 2019). For a DMU with input-output allocation (x_0, y_0) facing input prices v_0 and output prices w_0, the output oriented Farrell profit efficiency is defined as

$$PE = \sup_{\theta,x,y} \{\theta : w_0(\theta y_0) - v_0 x_0 \leq w_0 y - v_0 x, (x, y) \in \Psi \cap \mathcal{Z}\}, \tag{4}$$

where $\mathcal{Z}$ is a set imposing regularity conditions to ensure feasibility of the profit function (see more details in Färe et al., 2019). The profit efficiency can be decomposed as

$$PE = AE^{re} \times RE, \tag{5}$$

where AE^{re} is revenue-efficient allocative efficiency representing the improvement in profit due to the reallocation of inputs, RE is revenue efficiency defined as

$$RE(x_0, y_0, w_0) = \frac{RF(x_0, w_0)}{w_0 y_0}, \tag{6}$$

[1]See Sickles and Zelenyuk (2019) for more details and related discussion.

where $RF(x_0, w_0)$ is the classical revenue function defined as

$$RF(x_0, w_0) = \max_y \{w_0 y \mid (x_0, y) \in \Psi\}. \tag{7}$$

In practice, many inputs are relatively fixed in the short run or for a given year, especially for the case of hospitals that plan a budget in advance, hire personnel on relatively fixed (typically 12+ months) contracts, and put a lot of fixed inputs in (e.g., building, beds, equipment, etc.). Thus, revenue optimization becomes a natural special case of profit optimization and it is reasonable to look at revenue efficiency to analyse particular years or a short period of time. However, price information is not always available or if available, it might not represent market valuations. As a result, researchers often resort to analysing the technical efficiency part of revenue efficiency, and the most popular here is the Farrell-Debreu technical efficiency, which is a natural component in the decomposition of revenue efficiency (and thus of profit efficiency). The output oriented Farrell-Debreu technical efficiency for a DMU with input-output allocation (x_0, y_0) is defined as

$$\lambda(x_0, y_0) = \sup_\lambda \{\lambda \mid (x_0, \lambda y_0) \in \Psi\}. \tag{8}$$

Färe and Zelenyuk (2003) proposed an economic theory-based approach to obtain aggregate efficiencies from individual efficiencies discussed above. Simar and Zelenyuk (2007) (hereafter SZ2007) extended the result in Färe and Zelenyuk (2003) to the aggregation within and between sub-groups in a given group. Here, we focus our discussion on the aggregation of the technical efficiency.

To facilitate the discussion, let us assume we have a group of n DMUs with an input-output allocation set $\mathcal{X}_n = \{(X_i, Y_i) \mid i = 1, \ldots, n\}$, which come from L mutually exclusive and collectively exhaustive sub-groups (according to some exogenous economic criteria), and all DMUs face the same output prices, denoted from now on as w. Now let us denote the input-output allocation of each sub-group ℓ as $\mathcal{X}^\ell_{n_\ell} = \left\{\left(X^\ell_i, Y^\ell_i\right) \;\middle|\; i = 1, \ldots, n_\ell\right\}$, $\ell \in \{1, \cdots, L\}$. Following SZ2007, the sub-group aggregate technical efficiency for the sub-group ℓ can be obtained as [2]

$$\overline{TE}^\ell = \sum_{i=1}^{n_\ell} \lambda\left(X^\ell_i, Y^\ell_i\right) \times S^\ell_i, \quad S^\ell_i = \frac{w Y^\ell_i}{w \sum_{i=1}^{n_\ell} Y^\ell_i}, \quad \ell = 1, \cdots, L, \tag{9}$$

and the entire group aggregate technical efficiency can be obtained as

$$\overline{TE} = \sum_{\ell=1}^{L} \overline{TE}^\ell \times S^\ell, \quad S^\ell = \frac{w \sum_{i=1}^{n_\ell} Y^\ell_i}{w \sum_{\ell=1}^{L} \sum_{i=1}^{n_\ell} Y^\ell_i}. \tag{10}$$

[2] We use $\lambda\left(X^\ell_i, Y^\ell_i\right)$ and $\hat{\lambda}\left(X^\ell_i, Y^\ell_i \mid \mathcal{X}_n\right)$ (will be discussed later) to denote respectively the true and the estimate of true technical efficiency of a random point $\left(X^\ell_i, Y^\ell_i\right)$ in sub-group ℓ, and we replace $\left(X^\ell_i, Y^\ell_i\right)$ with $\left(x^\ell, y^\ell\right)$ to denote the estimated and true efficiencies of a fixed point $\left(x^\ell, y^\ell\right)$.

2.2 Envelopment Estimators

In practice, the true technical efficiency of each DMU is unknown, thus we need to obtain its estimate to proceed with the estimation of the aggregate efficiencies. There are many approaches in literature to estimate individual technical efficiencies, and here we focus on envelopment estimators.

If the constant returns to scale (CRS) assumption is imposed on the technology frontier, one can estimate technical efficiency for a random point $\left(X_i^\ell, Y_i^\ell\right)$ in sub-group ℓ using the CRS-DEA estimator [3], which is given by

$$\hat{\lambda}\left(X_i^\ell, Y_i^\ell \mid \mathcal{X}_n\right) \equiv \max_{\zeta_1,\ldots,\zeta_n,\lambda} \left\{ \lambda : \sum_{k=1}^{n} \zeta_k Y_k \geq \lambda Y_i^\ell, \sum_{k=1}^{n} \zeta_k X_k \leq X_i^\ell, \right. \\ \left. \lambda \geq 0, \forall \zeta_k \geq 0 \right\}. \tag{11}$$

Alternatively, if the variable returns to scale is imposed on the technology frontier, one can use the VRS-DEA estimator (Färe et al., 1983; Banker et al., 1984), which is given by

$$\hat{\lambda}\left(X_i^\ell, Y_i^\ell \mid \mathcal{X}_n\right) \equiv \max_{\zeta_1,\ldots,\zeta_n,\lambda} \left\{ \lambda : \sum_{k=1}^{n} \zeta_k Y_k \geq \lambda Y_i^\ell, \sum_{k=1}^{n} \zeta_k X_k \leq X_i^\ell, \right. \\ \left. \lambda \geq 0, \forall \zeta_k \geq 0, \sum_{k=1}^{n} \zeta_k = 1 \right\}. \tag{12}$$

Moreover, if the convexity assumption is not imposed, one can use the FDH estimator (Deprins et al., 1984) given by

$$\hat{\lambda}\left(X_i^\ell, Y_i^\ell \mid \mathcal{X}_n\right) \equiv \max_{\zeta_1,\ldots,\zeta_n,\lambda} \left\{ \lambda : \sum_{k=1}^{n} \zeta_k Y_k^\ell \geq \lambda Y_i^\ell, \sum_{k=1}^{n} \zeta_k X_k \leq X_i^\ell, \right. \\ \left. \lambda \geq 0, \zeta_k \in \{0,1\}, \sum_{k=1}^{n} \zeta_k = 1 \right\}. \tag{13}$$

Here individual efficiencies in all sub-groups are estimated using the same technology, allowing us to measure efficiencies across sub-groups of DMUs with respect to a common benchmark (to make them comparable).

Although statistical properties of DEA and FDH estimators at a fixed point have been well-established in the literature [4], their statistical properties at random points $\left(X_i^\ell, Y_i^\ell\right)$ have recently been developed by Kneip et al. (2015) (hereafter, KSW). Under

[3] This approach was largely initiated by Farrell (1957) and then generalized and popularized by Charnes et al. (1978), with many developments after.

[4] See details in Kneip et al. (1998), Park et al. (2000), Kneip et al. (2008), Park et al. (2010)

the assumption that technology and data generating processes (DGP) satisfy regularity conditions specified in KSW and as $n \to \infty$, the basic results established in KSW can be stated as follows

$$E\left[\hat{\lambda}\left(X_i^\ell, Y_i^\ell \mid \mathcal{X}_n\right) - \lambda\left(X_i^\ell, Y_i^\ell\right)\right] = C_0 n^{-\kappa} + R_{n,\kappa} \tag{14}$$

$$E\left[\left(\hat{\lambda}\left(X_i^\ell, Y_i^\ell \mid \mathcal{X}_n\right) - \lambda\left(X_i^\ell, Y_i^\ell\right)\right)^2\right] = o\left(n^{-\kappa}\right) \tag{15}$$

$$\left|COV\left[\hat{\lambda}\left(X_i^\ell, Y_i^\ell \mid \mathcal{X}_n\right) - \lambda\left(X_i^\ell, Y_i^\ell\right), \hat{\lambda}\left(X_j^\ell, Y_j^\ell \mid \mathcal{X}_n\right) - \lambda\left(X_j^\ell, Y_j^\ell\right)\right]\right| = o\left(n^{-1}\right) \tag{16}$$

Here, the values of the constant term C_0, the rate κ and the remainder term $R_{n,\kappa}$ depend on the dimension of input-output space and types of estimators used (see Table 1).

Table 1: Rate of convergence of envelopment estimators

Estimators	κ	$R_{n,\kappa}$
CRS-DEA	$2/(p+q)$	$O\left(n^{-3\kappa/2}\left(\log n\right)^{\alpha_1}\right)$
VRS-DEA	$2/(p+q+1)$	$O\left(n^{-3\kappa/2}\left(\log n\right)^{\alpha_2}\right)$
FDH	$1/(p+q)$	$O\left(n^{-2\kappa}\left(\log n\right)^{\alpha_3}\right)$

Notes:

* In all cases above, the values of $\alpha_j > 1,\ j = 1, 2, 3$ and other details are given in KSW.

For the aggregate technical efficiency, one can obtain its estimators by plugging in the envelopment estimators of individual technical efficiencies into equation (9). The asymptotic properties of aggregate technical efficiency have been recently developed by Simar and Zelenyuk (2018) (hereafter SZ2018), with further improvements for finite samples in Simar and Zelenyuk (2020) (hereafter SZ2020), and we will discuss them in the next section.

2.3 Confidence intervals for aggregate efficiency

2.3.1 Bootstrapping approach

The envelopment estimators of aggregate efficiency inherit the statistical properties from envelopment estimators of individual efficiency, which are biased and the bias is of a higher order than the variance when the dimension of the production model (i.e., $p+q$) increases (Kneip et al., 2015). As a result, the conventional statistical inferences fail to apply to the aggregate efficiency.

SZ2007 attempted to correct the bias and construct a confidence interval for the aggregate efficiency. They extended the subsample bootstrapping technique proposed in Kneip et al. (2008) to the context of the group-wise heterogeneity. Assuming that the DGP characterising the technology satisfies the assumptions specified in SZ2007, we can follow their bootstrap algorithm to obtain confidence intervals for the true aggregate technical efficiencies of the entire group, $\overline{TE}$, as well as of each sub-group, $\overline{TE}^{\ell},\ \ell \in \{1, \cdots, L\}$. The algorithm is summarised as follows

Step 1 Obtain estimates of $\overline{TE}$ and $\overline{TE}^{\ell},\ \ell \in \{1, \cdots, L\}$: Utilise envelopment estimators to obtain estimates of individual efficiency scores from the original sample on inputs and outputs for all DMUs, $\mathcal{S}_n = \{(x_i, y_i) \mid i = 1, \ldots, n\}$ [5]. Obtain estimates of $\overline{TE}$ and $\overline{TE}^{\ell}$ by aggregating individual efficiency scores using formulas (9) and (10), denote them as $\widehat{\overline{TE}}$ and $\widehat{\overline{TE}}^{\ell}$, respectively.

Step 2. Obtain b^{th} bootstrap samples: For each sub-group, ℓ, determine subsample size $s_\ell = \lfloor n_\ell^{\gamma_\ell} \rfloor$ ($\gamma_\ell \in (0, 1)$ and chosen by researchers). Generate a bootstrap sample for each sub-group ℓ, $\mathcal{S}^{*\ell}_{s_\ell,b} = \left\{ \left(x^{*\ell}_{j,b}, y^{*\ell}_{j,b}\right) \;\middle|\; j = 1, \ldots, s_\ell \right\}$, by resampling (uniformly, independently and with replacement) s_ℓ pairs out n_ℓ pairs $\left(x^{\ell}, y^{\ell}\right)$ from the original sample of each sub-group $\ell \in \{1, \cdots, L\}$. Do it separately for each sub-group $\ell \in \{1, \cdots, L\}$ and construct a pooled bootstrap sample denoted as $\mathcal{S}^*_{s,b} = \cup_{\ell=1}^{L} \mathcal{S}^{*\ell}_{s_\ell,b}$, $s = \sum_{\ell=1}^{L} s_\ell$

Step 3. Obtain b^{th} bootstrap estimates of $\overline{TE}$ and $\overline{TE}^{\ell},\ \ell \in \{1, \cdots, L\}$: Perform the same procedure described in Step 1 to compute bootstrap estimates of $\overline{TE}$ and $\overline{TE}^{\ell}$ but with respect to the frontier constructed from the pooled bootstrap sample $\mathcal{S}^*_{s,b}$, denote them as $\widehat{\overline{TE}}^{*}_b$ and $\widehat{\overline{TE}}^{*\ell}$, respectively.

Step 4. Repeat Step 2 and Step 3 B times. At the end, we obtain B bootstrap estimates of aggregate efficiencies for entire group $\left\{\widehat{\overline{TE}}^{*}_b\right\}_{b=1}^{B}$ and for each sub-group $\left\{\widehat{\overline{TE}}^{*\ell}_b\right\}_{b=1}^{B}$, $\ell \in \{1, \cdots, L\}$.

Step 5. Use $\left\{\widehat{\overline{TE}}^{*}_b\right\}_{b=1}^{B}$ and $\left\{\widehat{\overline{TE}}^{*\ell}_b\right\}_{b=1}^{B}$ with the formulas discussed below to construct the bootstrap-based confidence intervals for $\overline{TE}$ and $\overline{TE}^{\ell},\ \ell \in \{1, \cdots, L\}$, respectively.

[5] Here, we change notations to emphasize that estimates are obtained at fixed points.

The formula to construct $(1-\alpha)$ bootstrap confidence interval for the entire group aggregate technical efficiency, $\overline{TE}$, is

$$\widehat{\overline{TE}} + \hat{a}_\alpha \leq \overline{TE} \leq \widehat{\overline{TE}} + \hat{b}_\alpha, \tag{17}$$

where $\hat{b}_\alpha$ and $\hat{a}_\alpha$ are $(\alpha/2)-$quantile and $(1-\alpha/2)-$quantile of the set $\left\{\widehat{\overline{TE}}_b^* - \widehat{\overline{TE}}\right\}_{b=1}^{B}$, respectively. Similarly, the formula to construct $(1-\alpha)$ bootstrap confidence interval for each sub-group aggregate technical efficiency, $\overline{TE}^\ell,\ \ell \in \{1, \cdots, L\}$, is

$$\widehat{\overline{TE}}^\ell + \hat{a}_\alpha^\ell \leq \overline{TE}^\ell \leq \widehat{\overline{TE}}^\ell + \hat{b}_\alpha^\ell, \tag{18}$$

where $\hat{b}_\alpha^\ell$ and $\hat{a}_\alpha^\ell$ are $(\alpha/2)-$quantile and $(1-\alpha/2)-$quantile of the set $\left\{\widehat{\overline{TE}}_b^{*\ell} - \widehat{\overline{TE}}^\ell\right\}_{b=1}^{B}$, respectively. The interested reader can find formulas to calculate bootstrap estimates of bias and standard error for aggregate efficiencies in SZ2007.

The main complication of the subsampling bootstrap approach is that the choice of a subsample size (determined by γ_ℓ) is important in practice and the conclusions for empirical studies might be different quantitatively and even qualitatively for different subsample sizes.

2.3.2 Central Limit Theorem Approach

SZ2018 proposed a novel approach for interpreting the aggregate efficiencies as estimators of the ratio of two population means, which is useful for deriving asymptotic properties of these aggregate efficiencies. The goal of this section is to extend their approach to the context when there are several sub-groups in the population, as for those discussed above and earlier in SZ2007 for the bootstrap approach. Specifically, equation (9) can be rewritten as

$$\overline{TE}^\ell = \frac{(1/n_\ell)\sum_{i=1}^{n_\ell} wY_i^{\ell\partial}}{(1/n_\ell)\sum_{i=1}^{n_\ell} wY_i^\ell} = \frac{(1/n_\ell)\sum_{i=1}^{n_\ell} Z_i^{\ell\partial}}{(1/n_\ell)\sum_{i=1}^{n} Z_i^\ell}, \tag{19}$$

where, $Z_i^\ell \equiv wY_i^\ell$ is the revenue of DMU i and $Z_i^{\ell\partial} \equiv wY_i^{\ell\partial} = \lambda\left(X_i^\ell, Y_i^\ell\right) wY_i^\ell$ can be viewed as the monetary value (under prices w) of the projection of the observation Y_i^ℓ on the boundary of the technology set.

Adapting the notation of SZ2018 to the sub-group case, let $\left\{\left(Z_i^{\ell\partial}, Z_i^\ell\right) \mid i = 1, \ldots, n_\ell\right\}$ be a set of random realizations of a random vector

$$\begin{bmatrix} Z^{\ell\partial} \\ Z^\ell \end{bmatrix} = \begin{bmatrix} \lambda\left(X^\ell, Y^\ell\right) Z^\ell \\ Z^\ell \end{bmatrix}, \tag{20}$$

with its first two moments denoted as

$$\mu^{\ell} = \begin{bmatrix} E\left(Z^{\ell\partial}\right) \\ E\left(Z^{\ell}\right) \end{bmatrix} = \begin{bmatrix} \mu_1^{\ell} \\ \mu_2^{\ell} \end{bmatrix}, \tag{21}$$

$$\Sigma^{\ell} = \begin{bmatrix} VAR\left(Z^{\ell\partial}\right) & COV\left(Z^{\ell\partial}, Z^{\ell}\right) \\ COV\left(Z^{\ell\partial}, Z^{\ell}\right) & VAR\left(Z^{\ell}\right) \end{bmatrix} = \begin{bmatrix} \left(\sigma_1^{\ell}\right)^2 & \sigma_{12}^{\ell} \\ \sigma_{12}^{\ell} & \left(\sigma_2^{\ell}\right)^2 \end{bmatrix}. \tag{22}$$

As a result, the aggregate technical efficiency in equation (19) can be viewed as a natural estimator - the ratio of the sample means - for our parameter of interest $\tau^{\ell} = \dfrac{\mu_1^{\ell}}{\mu_2^{\ell}}$, i.e.,

$$\hat{\tau}^{\ell} = \frac{\hat{\mu}_1^{\ell}}{\hat{\mu}_2^{\ell}} = \frac{(1/n_{\ell}) \sum_{i=1}^{n_{\ell}} Z_i^{\ell\partial}}{(1/n_{\ell}) \sum_{i=1}^{n_{\ell}} Z_i^{\ell}}. \tag{23}$$

2.3.2.1 Central Limit Theorems

As pointed out by SZ2018, if $Z_i^{\ell\partial}$ is observable, one can use the conventional method to derive the asymptotic distribution for $\hat{\tau}^{\ell}$

$$\sqrt{n_{\ell}}\left(\hat{\tau}^{\ell} - \tau^{\ell}\right) \xrightarrow{d} N\left(0, \left(\sigma_{\tau}^{\ell}\right)^2\right), \tag{24}$$

where

$$\left(\sigma_{\tau}^{\ell}\right)^2 = n_{\ell}\left(\sigma_{\hat{\tau}^{\ell}}^{\ell}\right)^2 = \left(\tau^{\ell}\right)^2 \left(\frac{\left(\sigma_1^{\ell}\right)^2}{\left(\mu_1^{\ell}\right)^2} + \frac{\left(\sigma_2^{\ell}\right)^2}{\left(\mu_2^{\ell}\right)^2} - 2\frac{\sigma_{12}^{\ell}}{\mu_1^{\ell}\mu_2^{\ell}} \right). \tag{25}$$

In practice, we do not observe $Z_i^{\ell\partial}$ since the values of function $\lambda\left(X_i^{\ell}, Y_i^{\ell}\right)$ are unknown. However, we can obtain the estimates of $\lambda\left(X_i^{\ell}, Y_i^{\ell}\right)$ using nonparametric envelopment estimators discussed in the previous section. With those estimates, one can replace the unknowns in the numerator of equation (23) with their estimated values to obtain another estimator of τ^{ℓ}

$$\hat{\hat{\tau}}^{\ell} = \frac{\hat{\hat{\mu}}_1^{\ell}}{\hat{\mu}_2^{\ell}} = \frac{(1/n_{\ell}) \sum_{i=1}^{n_{\ell}} \hat{Z}_i^{\ell\partial}}{(1/n_{\ell}) \sum_{i=1}^{n_{\ell}} Z_i^{\ell}}, \tag{26}$$

where $\hat{Z}_i^{\ell\partial} = \hat{\lambda}\left(X_i^{\ell}, Y_i^{\ell}\right) w Y_i^{\ell}$.

The conventional central limit theorem, however, fails to apply to $\hat{\hat{\tau}}^{\ell}$ since envelopment estimators of $\lambda\left(X_i^{\ell}, Y_i^{\ell}\right)$ are biased and, in most instances, the bias does not converge to zero fast enough with the increase of sample size (e.g., it happens for CRS-DEA when $p + q > 3$, for VRS-DEA when $p + q > 2$ and for FDH when $p + q > 1$). SZ2018 have overcome the issue and developed new central limit theorems for the aggregate efficiency for the entire group by adapting and generalizing the results from KSW to correct bias and control convergence rates for both bias and variance. In the following, we will adapt the theorems developed in SZ2018 for the sub-group aggregate efficiency and theorems developed in KSW for sub-group mean efficiency with the assumption that $n_{\ell}/n \to c_{\ell}$

as $n \to \infty$, for any constant $c_\ell \in (0,1)$. To prepare the notations, let us denote the sub-group mean efficiency as $\overline{\lambda}^\ell$, i.e.,

$$\overline{\lambda}^\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \hat{\lambda}\left(X_i^\ell, Y_i^\ell | \mathcal{X}_n\right), \tag{27}$$

and the sub-group average efficiency in population as μ_λ^ℓ, i.e.,

$$\mu_\lambda^\ell = \mathrm{E}\left(\lambda\left(X^\ell, Y^\ell\right)\right), \tag{28}$$

and the variance of sub-group efficiency scores in the population as $\left(\sigma_\lambda^\ell\right)^2$, i.e.,

$$\left(\sigma_\lambda^\ell\right)^2 = \mathrm{Var}\left(\lambda\left(X^\ell, Y^\ell\right)\right). \tag{29}$$

The adapted results for the sub-group context can be summarised in the following two theorems.

Theorem 2.1. *Under the appropriate set of assumptions described in Theorem 3.1, 3.2 or 3.3 of KSW, for $p+q \leq 5$ if a CRS- DEA estimator is used and Ψ is CRS and is convex, for $p+q \leq 4$ if VRS-DEA estimator is used and Ψ is convex, for $p+q \leq 3$ if FDH estimator is used and satisfies free disposability of inputs and outputs, and if $n_\ell/n \to c_\ell$ as $n \to \infty$, when $n \to \infty$ we have*

$$\frac{\sqrt{n_\ell}}{\hat{\sigma}_\lambda^\ell}\left(\overline{\lambda}^\ell - \hat{B}^\ell_{\mu_\lambda^\ell,\kappa} - \mu_\lambda^\ell + R^\ell_{n,\kappa}\right) \xrightarrow{d} N(0,1), \tag{30}$$

and

$$\frac{\sqrt{n_\ell}}{\hat{\hat{\sigma}}_{\tau^\ell}^\ell}\left(\hat{\hat{\tau}}^\ell - \hat{B}^\ell_{\tau^\ell,\kappa} - \tau^\ell + R^\ell_{n,\kappa}\right) \xrightarrow{d} N(0,1), \tag{31}$$

where $R^\ell_{n,\kappa} = o\left(n^{-\kappa}\right)$, $\hat{B}^\ell_{\mu_\lambda^\ell,\kappa}$ and $\hat{B}^\ell_{\tau^\ell,\kappa}$ are respectively the generalized jackknife estimators of biases of $\overline{\lambda}^\ell$ and $\hat{\hat{\tau}}^\ell$, $\hat{\sigma}_\lambda^\ell$ is the empirical version of σ_λ^ℓ, and $\hat{\hat{\sigma}}_{\tau^\ell}^\ell$ is a consistent estimate of $\sigma_{\tau^\ell}^\ell$ given by

$$\left(\hat{\hat{\sigma}}_{\tau^\ell}^\ell\right)^2 = \hat{\hat{\tau}}^\ell \left[\frac{\left(\hat{\hat{\sigma}}_1^\ell\right)^2}{\left(\hat{\hat{\mu}}_1^\ell\right)^2} + \frac{\left(\hat{\sigma}_2^\ell\right)^2}{\left(\hat{\mu}_2^\ell\right)^2} - 2\frac{\hat{\hat{\sigma}}_{12}^\ell}{\hat{\hat{\mu}}_1^\ell \hat{\mu}_2^\ell}\right], \tag{32}$$

with $\hat{\hat{\sigma}}_1^\ell, \hat{\sigma}_2^\ell$, and $\hat{\hat{\sigma}}_{12}^\ell$ being empirical versions of $\sigma_1^\ell, \sigma_2^\ell$, and σ_{12}^ℓ, respectively.

Theorem 2.2. *Under the appropriate set of assumptions described in Theorem 3.1, 3.2 or 3.3 of KSW, for $p+q \geq 5$ if a CRS- DEA estimator is used and Ψ is CRS and is convex, for $p+q \geq 4$ if VRS-DEA estimator is used and Ψ is convex, for $p+q \geq 3$ if FDH estimator is used and satisfies free disposability of inputs and outputs, and if $n_\ell/n \to c_\ell$ as $n \to \infty$, when $n \to \infty$ we have*

$$\frac{\sqrt{n_{\ell,\kappa}}}{\hat{\sigma}_\lambda^\ell}\left(\overline{\lambda}_\kappa^\ell - \hat{B}^\ell_{\mu_\lambda^\ell,\kappa} - \mu_\lambda^\ell + R^\ell_{n,\kappa}\right) \xrightarrow{d} N(0,1), \tag{33}$$

and

$$\frac{\sqrt{n_{\ell,\kappa}}}{\hat{\hat{\sigma}}_{\tau^\ell}}\left(\hat{\hat{\tau}}_{\kappa}^{\ell}-\hat{B}_{\tau^\ell,\kappa}^{\ell}-\tau^{\ell}+R_{n,\kappa}^{\ell}\right)\xrightarrow{d}N\left(0,1\right) \tag{34}$$

where $\overline{\lambda}_{\kappa}^{\ell}$ *and* $\hat{\hat{\tau}}_{\kappa}^{\ell}$ *are respectively subsample versions of* $\overline{\lambda}^{\ell}$ *and* $\hat{\hat{\tau}}^{\ell}$*, in the sense that the averages are taken over a random subsample* $\mathcal{X}_{n_{\ell,\kappa}}^{\ell*}\subseteq\mathcal{X}_{n_\ell}^{\ell}$ *of size* $n_{\ell,\kappa}=\min\left(\lfloor n^{2\kappa}\rfloor,n_\ell\right)$*. Formally*

$$\overline{\lambda}_{\kappa}^{\ell}=n_{\ell,\kappa}^{-1}\sum_{\left\{i\middle|\left(X_i^\ell,Y_i^\ell\right)\in\mathcal{X}_{n_{\ell,\kappa}}^{\ell*}\right\}}\hat{\lambda}\left(X_i^{\ell},Y_i^{\ell}\middle|\mathcal{X}_n\right), \tag{35}$$

and

$$\hat{\hat{\tau}}_{\kappa}^{\ell}=\frac{n_{\ell,\kappa}^{-1}\sum_{\left\{i\middle|\left(X_i^\ell,Y_i^\ell\right)\in\mathcal{X}_{n_{\ell,\kappa}}^{\ell*}\right\}}\hat{Z}_i^{\ell\partial}\left(\mathcal{X}_n\right)}{n_{\ell,\kappa}^{-1}\sum_{\left\{i\middle|\left(X_i^\ell,Y_i^\ell\right)\in\mathcal{X}_{n_{\ell,\kappa}}^{\ell*}\right\}}Z_i^{\ell}}, \tag{36}$$

where $\hat{Z}_i^{\partial}\left(\mathcal{X}_n\right)=\hat{\lambda}\left(X_i^{\ell},Y_i^{\ell}\middle|\mathcal{X}_n\right)Z_i^{\ell}$

The above theorems are adaptations of Theorems 4.3 and 4.4 in KSW and Theorems 2 and 3 in SZ2018. Some elaborations are needed, particularly, to determine the order of the remainder term, $R_{n,\kappa}^{\ell}$. If one directly applies the theorems in KSW and SZ2018, one should estimate the efficiency scores of sub-group ℓ using the sub-group sample only (a sample of size n_ℓ). As such, the order of the remainder term should be $o\left(n_\ell^{-\kappa}\right)$. However, in our context, we have several sub-groups in the population and more importantly, efficiencies of individual DMUs in all sub-groups are measured using the common benchmark (to make them comparable), i.e., estimated using the entire sample (a sample of size n). As a result, the remainder term, $R_{n,\kappa}^{\ell}$, has the order of $o\left(n^{-\kappa}\right)$. To accommodate this, we assume that $n_\ell/n\to c_\ell$ as $n\to\infty$, for any constant $c_\ell\in(0,1)$, and thus when $n\to\infty$, we obtain the same results as in KSW and SZ. The subsample sizes $n_{\ell,\kappa}$ in Theorem 2.2 are also modified to reflect this fact.

Remark. As in KSW and SZ2018, with $p+q=5$ for CRS-DEA estimator, $p+q=4$ for VRS-DEA estimator, and $p+q=3$ for FDH estimator, both Theorem 2.1 and Theorem 2.2 are applicable because of the specific forms of the remainder term, but Theorem 2.2 is more preferable since the neglected term in Theorem 2.2 (the remainder term scaled by $\sqrt{n_{\ell,\kappa}}$) converges to zero faster than the neglected term in Theorem 2.1 (the remainder term scaled by $\sqrt{n_\ell}$) (see more related discussion in KSW and SZ2018).

2.3.2.2 Implementation of the Bias Correction

The generalized jackknife estimator of bias for the sub-group aggregate and mean efficiencies can be deployed by adapting the procedure discussed in KSW and SZ2018. Following KSW and SZ2018, we utilise the generalized jackknife technique to estimate the bias of

$\overline{\lambda}^{\ell}$ and $\hat{\bar{\mu}}_1^{\ell}$, and then estimate the bias $\hat{B}^{\ell}_{\tau^{\ell},\kappa}$ of $\hat{\tau}^{\ell}$ using the relationship in (26). The procedure is as follows.

First, let us randomly split the samples of each group $\ell \in \{1, \cdots, L\}$, into two parts with their sizes being $m_{\ell_1} = \lfloor n_\ell/2 \rfloor$ and $m_{\ell_2} = n_\ell - m_{\ell_1}$. Obviously, for each sub-group ℓ, the sample $\mathcal{X}^{\ell}_{n_\ell}$ is split evenly if n_ℓ is even or almost evenly if n_ℓ is odd, and there are $\binom{n_\ell}{m_{\ell_1}}$ possible ways to split the sample $\mathcal{X}^{\ell}_{n_\ell}$. Now let us consider a random split, say h, and denote a random subset of size m_{ℓ_1} of each $\mathcal{X}^{\ell}_{n_\ell}$ as $\mathcal{X}^{\ell(1)}_{m_{\ell_1},h}$, and let $\mathcal{X}^{\ell(2)}_{m_{\ell_2},h}$ be the set of remaining input-output pairs in each $\mathcal{X}^{\ell}_{n_\ell}$. The entire group sample is also split accordingly. Without loss of generality, let us construct two random subsets of the entire group as $\mathcal{X}^{1}_{m_1,h} = \cup_{\ell=1}^{L} \mathcal{X}^{\ell(1)}_{m_{\ell_1},h}$ and $\mathcal{X}^{2}_{m_2,h} = \cup_{\ell=1}^{L} \mathcal{X}^{\ell(2)}_{m_{\ell_2},h}$.[6]

For $j \in \{1, 2\}$, let

$$\overline{\lambda}^{\ell(j)}_h = \left(m_{\ell_j}\right)^{-1} \sum_{\left(X^{\ell}_i, Y^{\ell}_i\right) \in \mathcal{X}^{\ell(j)}_{m_{\ell_j},h}} \hat{\lambda}\left(X^{\ell}_i, Y^{\ell}_i | \mathcal{X}^{(j)}_{m_j,h}\right), \quad \ell = 1, \cdots, L, \tag{37}$$

and

$$\hat{\bar{\mu}}^{\ell(j)}_{1,h} = \left(m_{\ell_j}\right)^{-1} \sum_{\left(X^{\ell}_i, Y^{\ell}_i\right) \in \mathcal{X}^{\ell(j)}_{m_{\ell_j},h}} \hat{Z}^{\partial}_i\left(\mathcal{X}^{(j)}_{m_j,h}\right), \quad \ell = 1, \cdots, L. \tag{38}$$

Now, define

$$\overline{\lambda}^{\ell *}_h = \frac{1}{2}\left(\overline{\lambda}^{\ell(1)}_h + \overline{\lambda}^{\ell(2)}_h\right), \quad \ell = 1, \cdots, L, \tag{39}$$

and

$$\hat{\bar{\mu}}^{\ell *}_{1,h} = \frac{1}{2}\left(\hat{\bar{\mu}}^{\ell(1)}_{1,h} + \hat{\bar{\mu}}^{\ell(2)}_{1,h}\right), \quad \ell = 1, \cdots, L. \tag{40}$$

Then

$$(2^{\kappa} - 1)^{-1}\left(\overline{\lambda}^{\ell *}_h - \overline{\lambda}^{\ell}\right), \quad \ell = 1, \cdots, L, \tag{41}$$

and

$$(2^{\kappa} - 1)^{-1}\left(\hat{\bar{\mu}}^{\ell *}_{1,h} - \hat{\bar{\mu}}^{\ell}_1\right), \quad \ell = 1, \cdots, L, \tag{42}$$

provides estimators of the bias terms of $\overline{\lambda}^{\ell}$ and $\hat{\bar{\mu}}^{\ell}_1$, respectively. Repeating the operations above for $h = 1, \ldots, H$, with $H \ll \left(\binom{n_1}{n_1/2} \wedge \cdots \wedge \binom{n_L}{n_L/2}\right)$ and averaging (41) and (42), we obtain jackknife estimators of the bias terms of $\overline{\lambda}^{\ell}$ and $\hat{\bar{\mu}}^{\ell}_1$, of which variance is reduced by a factor of H^{-1} compared to (41) and (42), respectively

$$\hat{B}^{\ell}_{\mu^{\ell}_{\lambda},\kappa} = H^{-1} \sum_{h=1}^{H} (2^{\kappa} - 1)^{-1}\left(\overline{\lambda}^{\ell *}_h - \overline{\lambda}^{\ell}\right), \quad \ell = 1, \cdots, L, \tag{43}$$

and

$$\hat{B}^{\ell}_{\mu^{\ell}_{1},\kappa} = H^{-1} \sum_{h=1}^{H} (2^{\kappa} - 1)^{-1}\left(\hat{\bar{\mu}}^{\ell *}_{1,h} - \hat{\bar{\mu}}^{\ell}_1\right), \quad \ell = 1, \cdots, L. \tag{44}$$

[6] We have 2^L different ways to construct 2 subsets of entire group from 2 subsets of L sub-groups.

The jackknife estimator of bias of $\hat{\hat{\tau}}_\ell$ is then given by

$$\hat{B}^{\ell}_{\tau^\ell,\kappa} = \frac{H^{-1}\sum_{h=1}^{H}\left(2^\kappa-1\right)^{-1}\left(\hat{\hat{\mu}}^{\ell*}_{1,h}-\hat{\hat{\mu}}^{\ell}_{1}\right)}{\hat{\mu}^{\ell}_{2}}, \quad \ell=1,\cdots,L. \tag{45}$$

2.3.2.3 Implementation of the Variance Correction

KSW and SZ2018 suggest to use empirical estimators of true variance of individual and aggregate efficiency, respectively, to make practical inference using the new central limit theorems developed therein. They proved that the empirical estimators of the variances are consistent estimators and the central limit theorems are still applicable when these estimators are used in place of the true values. However, as discussed in SZ2020, these estimators are biased by construction. Specifically, for the case of individual efficiency, a well-known result is

$$1 \leq \hat{\lambda}\left(X,Y|\mathcal{X}_n\right) \leq \lambda\left(X,Y\right), \tag{46}$$

with probability one. This result implies that the empirical version of variance of individual efficiency underestimates the population variance. The same logic is applied to the case of the aggregate efficiency since the biases in estimating aggregate efficiency come from individual efficiency estimators. The biases of the variances might reduce the accuracy of statistical inference using the new central limit theorems in finite samples.

To improve finite sample approximation by the central limit theorems, SZ2020 propose using the bias corrected versions for the estimators of variance of individual and aggregate efficiencies. Adapting SZ2020, the bias-corrected estimators of variance for sub-group mean and aggregate efficiency are given respectively by

$$\tilde{\sigma}^{\ell}_{\lambda} = \hat{\sigma}^{\ell}_{\lambda} + \left(\hat{B}^{\ell}_{\mu^\ell_\lambda,\kappa}\right)^2, \quad \ell=1,\cdots,L, \tag{47}$$

and

$$\tilde{\sigma}^{\ell}_{\tau^\ell} = \hat{\hat{\tau}}^{\ell}\left[\frac{\left(\tilde{\sigma}^{\ell}_{1}\right)^2}{\left(\hat{\hat{\mu}}^{\ell}_{1}\right)^2}+\frac{\left(\hat{\sigma}^{\ell}_{2}\right)^2}{\left(\hat{\mu}^{\ell}_{2}\right)^2}-2\frac{\hat{\hat{\sigma}}^{\ell}_{12}}{\hat{\hat{\mu}}^{\ell}_{1}\hat{\mu}^{\ell}_{2}}\right], \quad \ell=1,\cdots,L, \tag{48}$$

where

$$\tilde{\sigma}^{\ell}_{1} = \hat{\hat{\sigma}}^{\ell}_{1} + \left(\hat{B}^{\ell}_{\mu^\ell_1,\kappa}\right)^2, \quad \ell=1,\cdots,L. \tag{49}$$

As in SZ2020, Theorem 2.1 and Theorem 2.2 also apply if $\tilde{\sigma}^{\ell}_{\lambda}$ and $\tilde{\sigma}^{\ell}_{\tau^\ell}$ are used in place of $\hat{\sigma}^{\ell}_{\lambda}$ and $\hat{\hat{\sigma}}^{\ell}_{\tau^\ell}$, respectively.

2.3.2.4 Confidence Intervals

The $(1-\alpha)$ confidence interval for the sub-group mean and aggregate efficiency can be constructed respectively as

$$\left[\overline{\lambda}^{\ell} - \hat{B}^{\ell}_{\mu^{\ell}_{\lambda},\kappa} \pm z_{1-\alpha/2}\frac{\hat{\sigma}^{\ell}_{\lambda}}{\sqrt{n_{\ell}}}\right], \quad \ell = 1, \cdots, L, \tag{50}$$

and

$$\left[\hat{\overline{\tau}}^{\ell} - \hat{B}^{\ell}_{\tau^{\ell},\kappa} \pm z_{1-\alpha/2}\frac{\hat{\hat{\sigma}}^{\ell}_{\tau^{\ell}}}{\sqrt{n_{\ell}}}\right], \quad \ell = 1, \cdots, L, \tag{51}$$

when Theorem 2.1 can be applied (here $z_{1-\alpha/2}$ is the corresponding quantile of the standard normal distribution), and

$$\left[\overline{\lambda}^{\ell}_{\kappa} - \hat{B}^{\ell}_{\mu^{\ell}_{\lambda},\kappa} \pm z_{1-\alpha/2}\frac{\hat{\sigma}^{\ell}_{\lambda}}{\sqrt{n_{\ell,\kappa}}}\right], \quad \ell = 1, \cdots, L, \tag{52}$$

and

$$\left[\hat{\overline{\tau}}^{\ell}_{\kappa} - \hat{B}^{\ell}_{\tau^{\ell},\kappa} \pm z_{1-\alpha/2}\frac{\hat{\hat{\sigma}}^{\ell}_{\tau^{\ell}}}{\sqrt{n_{\ell,\kappa}}}\right], \quad \ell = 1, \cdots, L, \tag{53}$$

when Theorem 2.2 can be applied.

Following SZ2020, we can use the bias corrected version for the estimators of variance to obtain the improved estimates of the $(1-\alpha)$ confidence interval for sub-group mean and aggregate efficiency respectively as

$$\left[\overline{\lambda}^{\ell} - \hat{B}^{\ell}_{\mu^{\ell}_{\lambda},\kappa} \pm z_{1-\alpha/2}\frac{\tilde{\sigma}^{\ell}_{\lambda}}{\sqrt{n_{\ell}}}\right], \quad \ell = 1, \cdots, L, \tag{54}$$

and

$$\left[\hat{\overline{\tau}}^{\ell} - \hat{B}^{\ell}_{\tau^{\ell},\kappa} \pm z_{1-\alpha/2}\frac{\tilde{\hat{\sigma}}^{\ell}_{\tau^{\ell}}}{\sqrt{n_{\ell}}}\right], \quad \ell = 1, \cdots, L, \tag{55}$$

when Theorem 2.1 can be applied, and

$$\left[\overline{\lambda}^{\ell}_{\kappa} - \hat{B}^{\ell}_{\mu^{\ell}_{\lambda},\kappa} \pm z_{1-\alpha/2}\frac{\tilde{\sigma}^{\ell}_{\lambda}}{\sqrt{n_{\ell,\kappa}}}\right], \quad \ell = 1, \cdots, L, \tag{56}$$

and

$$\left[\hat{\overline{\tau}}^{\ell}_{\kappa} - \hat{B}^{\ell}_{\tau^{\ell},\kappa} \pm z_{1-\alpha/2}\frac{\tilde{\hat{\sigma}}^{\ell}_{\tau^{\ell}}}{\sqrt{n_{\ell,\kappa}}}\right], \quad \ell = 1, \cdots, L, \tag{57}$$

when Theorem 2.2 can be applied.[7]

In their study, SZ2020 present results from many Monte-Carlo experiments showing that the improved estimates of confidence intervals persistently have better coverage, thus better accuracy and more reliability, especially in the cases of relatively large dimensions or a relatively small sample size. As a result, we will rely more on the improved confidence intervals in this study, although we will present results from both approaches.

[7]For the entire group confidence intervals, one can directly apply the relevant formulas in KSW, SZ2018 and SZ2020.

3 Data and Variables

3.1 Sample

In this study, we use annual data from public hospitals in Queensland, Australia in the period from FY 2012/13 to FY 2016/17 (the period after the National Health Reform Agreement).[8] Australia has a universal healthcare with a mix of private and public providers, and public hospitals are unarguably the most important institutions in the sector. Public hospital services in Australia are available free of charge for all persons who are eligible for Medicare and can access public hospitals as a public patient. Based on the data in FY 2016/17, public hospitals accounted for two-thirds of total hospital beds, three-fifths of total hospitalisations, and 30% of total healthcare expenditure in Australia.

The data is provided by Queensland Health, from two data collections namely Financial and Residential Activity Collection (FRAC) and Monthly Activity Collection (MAC). We obtain additional information, such as hospital peer groups and geographic location, from the Australian Institute of Health and Welfare (AIHW, 2015). Our sample includes only public acute hospitals.[9] Moreover, we only include hospitals that have complete data on all inputs and outputs for the entire study period, except for the case of the Gold Coast Hospital and the Gold Coast University Hospital, where we combine their data together.[10] Our final dataset contains 520 observations including 104 public acute hospitals over a five-year period.

We use three different criteria to classify public hospitals in Queensland into different sub-groups, which are: (i) Teaching status, (ii) Geographical location and (iii) Hospital size. For teaching status, hospitals in our sample are classified into two sub-groups which are non-teaching hospitals and teaching hospitals.[11] We obtain information about teaching status from FRAC. Our sample includes 29 teaching hospitals in FY 2016-17 and 118 observations with teaching status over the five-year period.[12]

For geographical location, hospitals are classified into non-remote sub-group (located

[8]FY stands for Financial Year, which in Australia starts on 1 July and ends on 30 June of the next calendar year.

[9]Public hospitals in Queensland include acute hospitals, women's and children's hospitals, and psychiatric hospitals. Women's and children's hospitals and psychiatric hospitals can be viewed as specialized hospitals (i.e., providing healthcare services to a specific target population or group of conditions).

[10]In 2013, the Gold Coast Hospital closed, and the Gold Coast University Hospital subsequently opened to replace it (all patients treated at the Gold Coast Hospital were transferred to the Gold Coast University Hospital).

[11]A hospital is defined as a teaching hospital if it affiliates with universities to provide undergraduate medical education as advised by the relevant state health authority.

[12]There are some hospitals changing teaching status during the studied period.

in major cities, inner regional and outer regional areas) and remote sub-group (located in remote and very remote areas). We obtain information about the remoteness of hospitals from AIHW (2015) in which remoteness of a hospital is measured by the physical road distance to its nearest urban center. The non-remote sub-group has 375 observations including 75 hospitals over five-year period. The remote sub-group has 145 observations including 29 hospitals over the five-year period.

For hospital size, we have two sub-groups which are a sub-group of small hospitals and a sub-group of large hospitals. Hospitals are classified as small or large based on their hospital peer groups developed by AIHW (2015). Specifically, the large sub-group is composed of principal referral hospitals, public acute group A hospitals, public acute group B hospitals, while the small sub-group includes public acute group C hospitals and public acute group D hospitals.[13] The large sub-group has 120 observations including 24 hospitals over the five-year period. The small sub-group has 400 observations including 80 hospitals over the five-year period.

3.2 Variables

In this study, we use three inputs and two outputs to model the production process of hospitals. The three inputs are: (i) Labour input, (ii) Capital input, and (iii) Consumable input. The two outputs are: (i) Outpatient output and (ii) Inpatient output. Each input and output will be discussed in detail in the below sections.

3.2.1 Input

3.2.1.1 Labour Input

Labour is the most important input of the hospital production process and has been incorporated into almost all studies in the literature on hospital efficiency (O'Neill et al., 2008). Following the common practice in the literature (e.g., Hao & Pegels, 1994; Burgess & Wilson, 1996; Magnussen, 1996; Harris et al., 2000; Grosskopf et al., 2001; Berta et al., 2010; Ferrier & Trivitt, 2013; Nayar et al., 2013; Chowdhury & Zelenyuk, 2016), we utilize full-time equivalent (FTE) staff as a proxy for labour input in our model.

Data on hospital staff is composed of six main categories which are salaried medical officers (MEO), nurses (NUR), diagnostic and health professionals (DHP), other personal care staff (OPCS), administrative and clerical staff (ACS), and domestic and other staff

[13]The five peer groups are listed in descending order of service diversification and volume of activities as follows: principal referral, acute group A, acute group B, acute group C, and acute group D. As indicated in AIHW (2015), the last two are usually smaller than the first three.

(DOS). Ideally, these six categories of personnel should be included in DEA models as separate inputs since personnel in different categories have different duties and contribute differently to the performance of hospitals (e.g. the first four categories directly perform clinical related activities while the duties of the last two are mainly administrative and clerical). However, as discussed in Section 2.2, when the dimension of input-output space increases, the convergence rates of envelopment estimators decrease significantly, reducing the reliability of analysis. To remedy the issue, observing that the six labour categories are highly correlated (see Table 2), we follow Daraio and Simar's (2007) approach (the approach based on Principal Component Analysis), to aggregate them into a single measure of labour input, called 'labour factor' (FLABOUR). Specifically, FLABOUR is constructed as follows

$$FLABOUR = 0.30MEO + 0.84TNUR + 0.24DHP + 0.05OPCS + 0.29ACS + 0.22DOS \quad (58)$$

The constructed 'labour factor' explains 98.79% of total inertia of the original data on different types of labour and has high correlations with each of the six labour categories (see Table 2). As a result, the dimension of production space is reduced without losing much information, and the 'labour factor' is a good representative of hospital labour input.

Table 2: Correlation matrix of labour inputs

	MEO	TNUR	DHP	OPCS	ACS	DOS	FLABOUR
MEO	1.00						
TNUR	1.00	1.00					
DHP	0.97	0.98	1.00				
OPCS	0.95	0.96	0.96	1.00			
ACS	0.98	0.99	0.98	0.95	1.00		
DOS	0.91	0.91	0.88	0.84	0.92	1.00	
FLABOUR	1.00	1.00	0.98	0.96	0.99	0.92	1.00

3.2.2 Capital input

The most appropriate measure of capital input for studying hospital efficiency is argued to be the utilization of capital in the production process (Worthington, 2004). Measurements of capital utilization are, however, usually not available in practice. As a result, researchers often utilize alternative measures that are proportional to the capital usage. The number

of beds is such a measure and is widely-used in hospital efficiency studies (see the review in O'Neill et al., 2008). Following the common practice in the literature, we use the number of beds (BEDS) as a proxy for the utilization of capital in this study. The data for the number of beds in Queensland is recorded at the end of each financial year and includes both the number of available beds and bed alternatives.

3.2.3 Consumable input

Following Biørn et al. (2003), Productivity Commission (PC, 2010), Chua et al. (2011), Besstremyannaya (2013), Chowdhury and Zelenyuk (2016), expenditures on drug, surgical and medical supplies (DMSUP) are used to represent hospital consumable input. The drug, surgical and medical supply expenditures are recorded at the current price at the end of each financial year, thus we utilize the consumer price index obtained from the Australian Bureau of Statistics to convert them to the constant price (using the year 2013/14 as the base year)

3.2.4 Outputs

3.2.4.1 Outpatient and Inpatient Outputs

Following the common practice in the literature (e.g., Zuckerman et al., 1994; Magnussen, 1996; Harris et al., 2000; Nayar & Ozcan, 2008; Nayar et al., 2013; Chowdhury & Zelenyuk, 2016), in this study, outpatient outputs of Queensland public hospitals (OUT) are measured by the number of non-admitted occasions of service including both emergency and non-emergency services for non-admitted patients.

For inpatient service quantities, it is not sufficient to measure the output by the raw counts of admitted patient episodes because of the need to distinguish inpatient care services based on their complexity and resources required. In our data, admitted patient episodes are categorised into more than 700 Diagnosis-Related Groups (DRGs), grouping together patients with similar diagnoses who require similar hospital services. In principle, each DRG represents a separate output, yet they need to be aggregated to make the model feasible. Following Burgess and Wilson (1996, 1998), Hofmarcher et al. (2002), Biørn et al. (2003), Clement et al. (2008), Nayar and Ozcan (2008), we use casemix weights to aggregate admitted patient episodes into a single measure of inpatient output, called casemix weighted inpatient episode (WEPS). Specifically, we multiply the number of inpatient episodes in each DRG by the DRG cost weight, then sum them up to get the number of casemix weighted inpatient episodes.[14] As discussed and illustrated

[14]We use the constant inlier cost weights of the year 2013/14 obtained from the Independent Hospital

by extensive Monte Carlo experiments in Zelenyuk (2020), this price-based aggregation approach in DEA often leads to some aggregation biases, but the biases are bounded by allocative efficiency and expected to be small and justifiable for practical reasons.

3.2.4.2 Output Prices

In order to obtain the technical aggregate efficiency, we need information about the relative price of inpatient output to outpatient output. For inpatient output, the price can be determined based on the Activity Based Funding models in which hospitals are paid one national efficient price (NEP) for each WEPS.[15] For outpatient output, we need an assumption to derive its price due to data limitation. As discussed above, outpatient output in our study is measured by the raw counts of non-admitted occasions of service. To derive the price for one raw count, we assume that the compositions of non-admitted occasions of service are the same across all hospitals in Queensland and utilise industry level data to calculate the expected price. With this assumption, the expected price of one non-admitted occasion of service is 0.09 NEP. As a result, the expected relative price of inpatient output to outpatient output is 1/0.09.

The descriptive statistics of all inputs and outputs are provided in Table 3 and their histograms are shown in Figure 1. We can see from Figure 1 that hospital inputs and outputs show substantial positive skewness, but this is typical for any healthcare data.

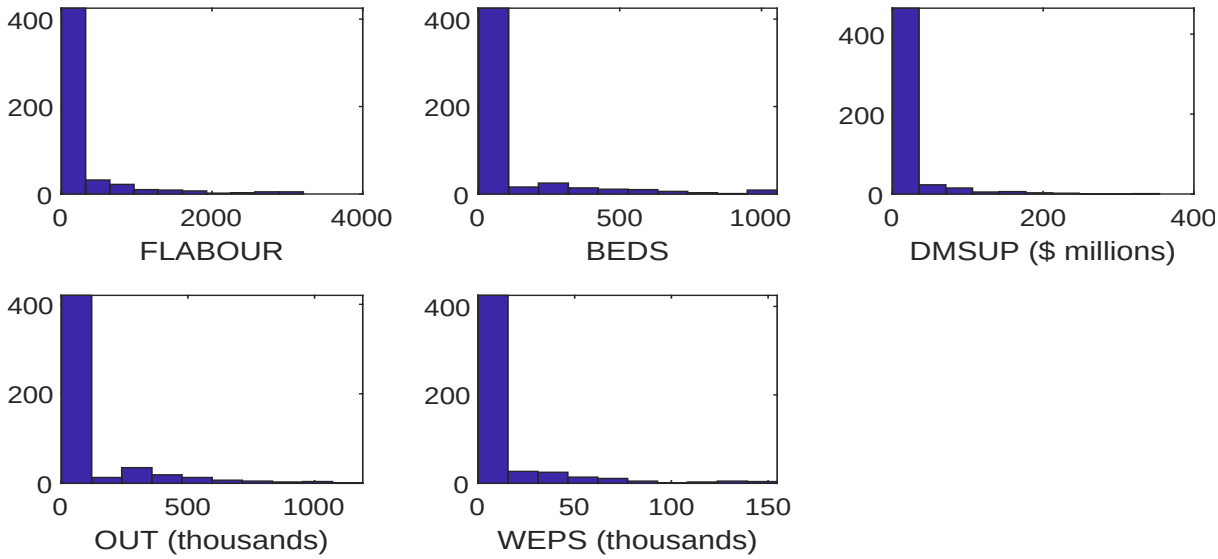

Figure 1: Histograms of hospital inputs and outputs.

Pricing Authority (IHPA, 2013).

[15]The NEP is determined by the Independent Hospital Pricing Authority for public hospital services through the analysis of data on actual activity and costs in public hospitals.

Table 3: The descriptive statistics of all variables used in this study

Variables	Description	Mean	Std. Dev.	Min	Max
Input variables					
FLABOUR	Labour aggregation using Daraio and Simar's (2007) approach	246.82	549.42	3.20	3211.24
BEDS	Total beds	99.67	196.48	3.00	1055.00
DMSUP	Drug and medical supply expenditure (2013/14 constant price) ($1,000,000s)	13.09	35.87	0.03	354.24
Output variables					
OUT	Non-admitted occasions of service (1,000s)	100.95	191.88	1.25	1190.22
WEPS	Casemix weighted inpatient episodes(1,000s)	11.47	25.78	0.03	154.66

4 Results and discussion

4.1 Individual efficiency

In this study, we obtain both the output oriented CRS-DEA and VRS-DEA estimates of individual efficiencies with respect to the grand frontier (i.e., pooling data over 5 years). In Figure 2 and Figure 3, we present the boxplots of estimated efficiency scores across years. For both estimators, we can see that the distribution of estimated efficiency scores are quite similar across years and in each year there are from 5 to 6 outliers (hospitals that are very inefficient relative to the majority of hospitals) which are mainly small and non-teaching hospitals. These outliers might face a different production environment (possibly more disadvantageous) compared to the others in our sample and deserve to be studied separately. We decided to not include these outliers and follow Zelenyuk and Zheka (2006) to trim 5% of outliers in the right tail of the estimated efficiency distributions. This results in a trimmed sample of 494 observations.[16]

As a sensitivity analysis, we compare the estimated densities of efficiency distributions

[16]The trimming is based on the distribution of estimated efficiency scores using the CRS-DEA estimator, but it is very similar to the case of VRS-DEA in terms of the number of and the composition of trimmed observations.

for the sample before and after trimming, as well as test for the hypothesis of equality of the two efficiency distributions using the adapted Li test (Simar & Zelenyuk, 2006). The estimated densities show that the data trimming does not significantly influence the distributions of efficiency scores (see Figures 4 and 5) and it is confirmed by the results of the adapted Li test.[17]

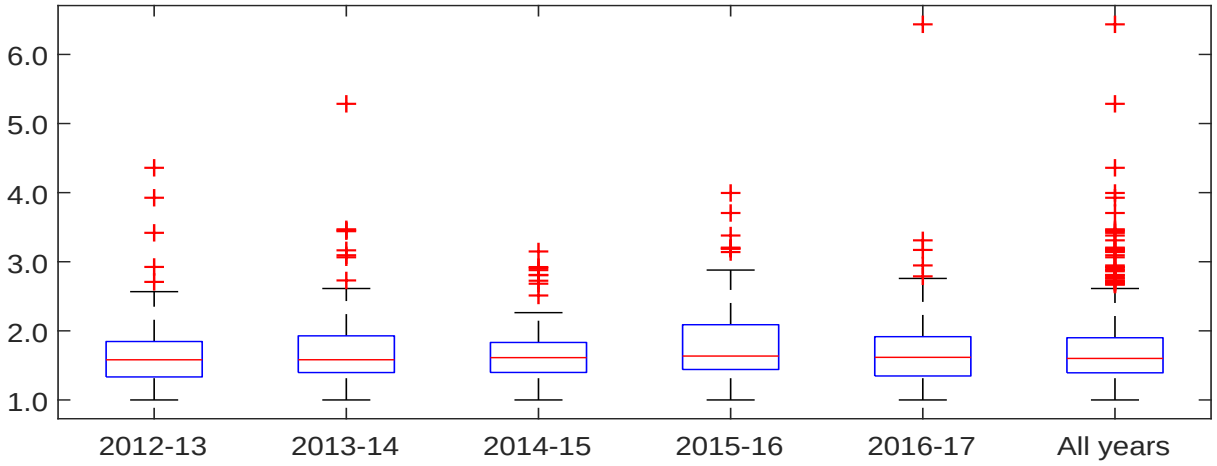

Figure 2: Boxplots of the estimated efficiency scores using the CRS-DEA estimator for different years.

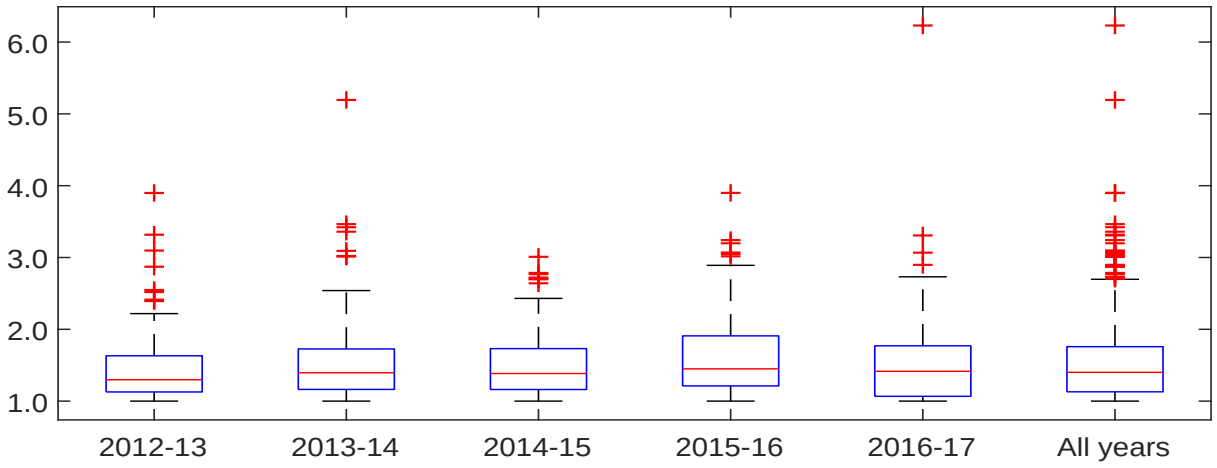

Figure 3: Boxplots of the estimated efficiency scores using the VRS-DEA estimator for different years.

[17]The p-values of the adapted Li test for the null hypotheses of equality of distributions of efficiency scores before and after trimming the data for CRS-DEA and VRS-DEA cases are 0.16 and 0.35, respectively, thus we do not reject the null hypotheses.

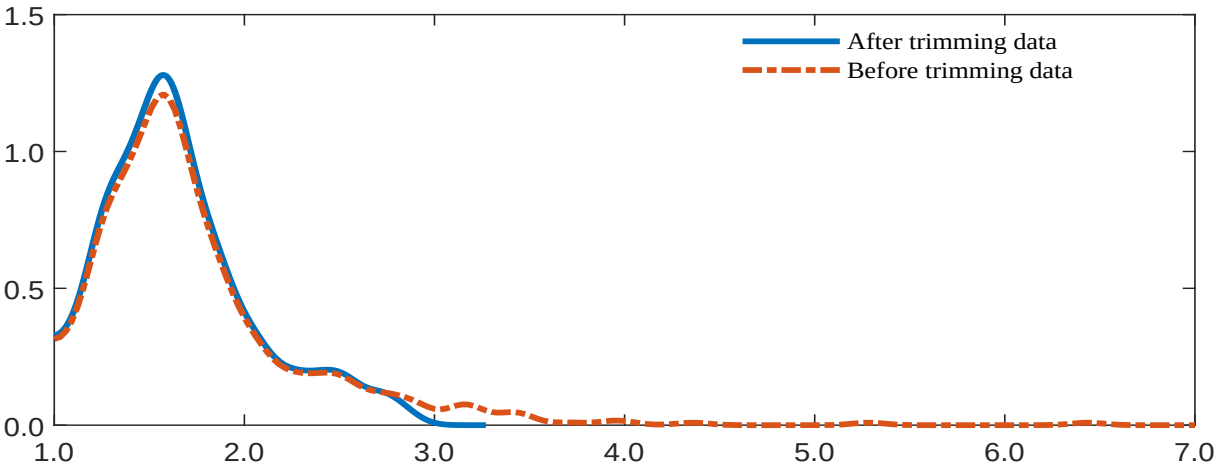

Figure 4: Density estimates of the estimated efficiency scores using the CRS-DEA estimator: Pre- vs. Post-trimming sample. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

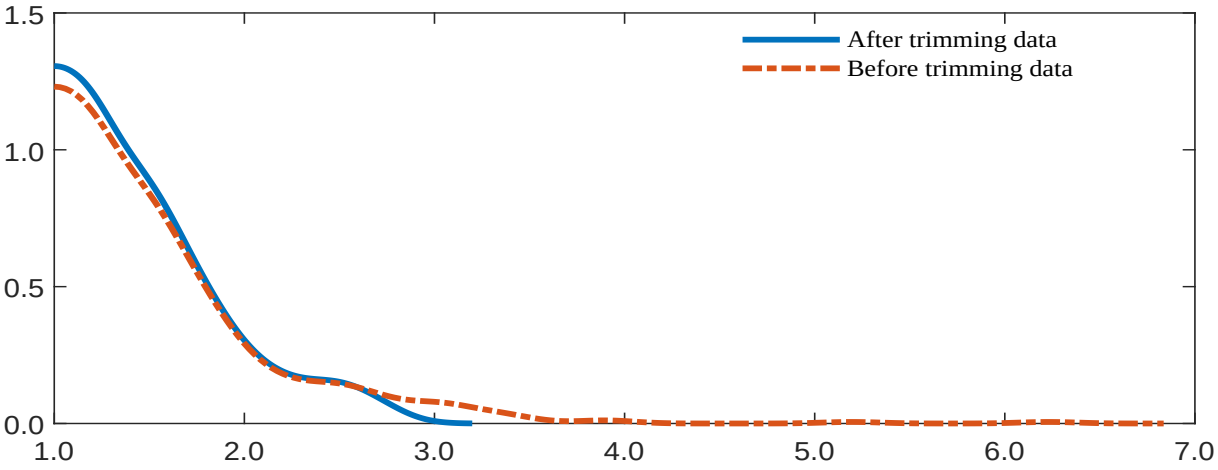

Figure 5: Density estimates of the estimated efficiency scores using the VRS-DEA estimator: Pre- vs. Post-trimming sample. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

4.2 Entire group efficiency

Before going to discussions about confidence intervals obtained using bootstrap and central limit theorem approaches, it is worth recalling that the choice of subsample sizes (determined by γ) is important in applied works, and the optimal choice of γ is still a largely unresolved question in the literature.[18] We propose here a method to choose γ, guided by the new central limit theorems. The optimal γ is chosen such that the resulted bootstrap standard deviation of aggregate efficiency estimators are equal to its asymptotic standard deviation (biased corrected) derived using the central limit theorems. Interestingly, the optimal γ determined by our method, $\gamma = 0.65$, happened to be very close to the rule-of-thumb γ, $\gamma = 0.7$, chosen in SZ2007 in their simulations and empirical illustration.[19]

4.2.1 Results for CRS-DEA

In this section and the following sections, the results reported include the DEA point estimates of aggregate and simple mean efficiencies, their bias corrected versions and the estimated standard deviations (using both bootstrap and central limit theorem approaches), the bias corrected standard deviations (using central limit theorem approach), and the corresponding lower bounds and upper bounds of confidence intervals using the bootstrap approach (SZ2007) and the central limit theorem approaches (KSW, SZ2018, SZ2020).

Table 4 shows the results for the case of the CRS-DEA estimator. Looking at Table 4, few remarks are in order. First, the CRS-DEA point estimates of aggregate and mean efficiencies are very similar, at 1.6 (i.e., efficiency level of 0.62) and 1.65 (i.e., efficiency level of 0.61), respectively. Second, the bias correction is substantially large for both aggregate and mean efficiencies, and for aggregate efficiency the bias correction is substantially different between the bootstrap and central limit theorem approaches, e.g., aggregate efficiency is corrected from 1.6 to 2.01 using the bootstrap approach but from 1.6 to 2.16 using the central limit theorem approach. Third, it turns out that under the central limit theorem approach biased corrected aggregate efficiency is higher (i.e., showing a lower average efficiency level) than biased corrected mean efficiency, indicating that some

[18] Simar and Wilson (2011) adapted the data-driven approach proposed in Politis et al. (2001) and Bickel and Sakov (2008) to select a subsample size in subsampling bootstrap for envelopment estimators and demonstrated its good performance by Monte Carlo evidence (see more discussions in Sickles & Zelenyuk, 2019).

[19] The optimal γ is tuned using a CRS-DEA model and will also be used for the sub-group analysis in the next section.

hospitals that are given larger weights (i.e., more economically important) in calculating aggregate efficiency are less efficient than the average when benchmarking toward the CRS frontier – the frontier that identifies the best practice socially-optimal scale.

The 95% confidence intervals for the entire group aggregate efficiency are also different between the two approaches. For the bootstrap approach, the confidence interval is from 1.62 to 2.13 (i.e., efficiency levels from 0.47 to 0.62), meanwhile for the central limit theorem approach (SZ2020) the confidence interval is from 1.87 to 2.39 (i.e., efficiency levels from 0.42 to 0.52).

Table 4: CRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Entire Group

		Aggregate Efficiency	Sample Mean
Bootstrap	DEA estimate	1.60	1.65
	Bias corrected	2.01	1.98
	Est. Std	0.13	0.07
	LB of est. CI	1.62	1.83
	UB of est. CI	2.13	2.10
CLT	DEA estimate	1.60	1.65
	Bias corrected	2.16	1.98
	Est. Std	0.12	0.03
	Bias corrected Std	0.13	0.04
	LB est. CI	1.89	1.90
	UB of est. CI	2.38	2.03
	LB of est. CI-Improved	1.87	1.88
	UB of est. CI-Improved	2.39	2.05

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

4.2.2 Results for VRS-DEA

The first important point to discuss when looking at the results using the VRS-DEA estimator is that both point estimates of aggregate and mean efficiencies and their bias corrected versions are substantially lower than their counterparts using the CRS-DEA estimator, especially for aggregate efficiency, e.g., the VRS-DEA estimate of aggregate efficiency is 1.13 (i.e., efficiency level of 0.88) and its biased corrected version (using central limit theorem approach) is 1.25 (efficiency level of 0.80). Moreover, for the VRS-

DEA estimator, the aggregate efficiency is significantly lower (i.e., showing a higher level of average efficiency) than the mean efficiency. Using the bootstrap approach, the 95% confidence interval of aggregate efficiency, which ranges from 1.18 to 1.24 (i.e., efficiency levels from 0.81 to 0.85), is on the left of and does not overlap with the the 95% confidence interval of the mean efficiency, which ranges from 1.62 to 1.81 (i.e., efficiency levels from 0.55 to 0.62). Similarly, with the central limit theorem approach, the 95% confidence interval of aggregate efficiency, which ranges from 1.00 to 1.41 (i.e., efficiency levels from 0.71 to 1.00), is on the left of and does not overlap with the 95% confidence interval of mean efficiency, which ranges from 1.58 to 1.84 (i.e., efficiency levels from 0.54 to 0.63).

The results discussed in this section and the previous section indicate that there exist substantial scale inefficiencies in the industry and particularly for the hospitals that are economically more important (i.e., hospitals receive high weights in the calculation of aggregate efficiency due to attaining higher revenue shares, as computed in (9)). To further investigate this, we examine the scale efficiencies of observations that are in the top 5% highest weights, those turn out to be very large hospitals with the number of beds being more than 434 beds - the 0.925-quantile of the number of beds in our sample. The result shows that all of these observations exhibit diseconomies of scale with the scale inefficiency level ranging from 0.21 to 0.47.[20]

[20]The scale efficiency score is computed as a ratio of the CRS-DEA technical efficiency score to the VRS-DEA technical efficiency score. The scale inefficiency level is obtained by subtracting the reciprocal of the scale efficiency score from one.

Table 5: VRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Entire Group

		Aggregate Efficiency	Sample Mean
Bootstrap	DEA estimate	1.13	1.46
	Bias corrected	1.21	1.73
	Est. Std	0.01	0.05
	LB of est. CI	1.18	1.62
	UB of est. CI	1.24	1.81
CLT	DEA estimate	1.13	1.46
	Bias corrected	1.25	1.76
	Est. Std	0.10	0.05
	Bias corrected Std	0.10	0.07
	LB est. CI	1.00	1.61
	UB of est. CI	1.41	1.82
	LB of est. CI-Improved	1.00	1.58
	UB of est. CI-Improved	1.41	1.84

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

4.3 Sub-group Efficiency by Teaching Status

In this section we investigate the sub-group efficiency of teaching and non-teaching hospitals in the industry.[21] As in the case of entire group efficiency, we also look at both CRS-DEA and VRS-DEA estimators.

4.3.1 Results for CRS-DEA

To compare the efficiency of teaching and non-teaching hospitals, we first estimate and visualize the density of estimated efficiency scores for each sub-group (Figure 6). The estimated density of the teaching sub-group is below the density of the non-teaching sub-group in the region of low efficiency scores (i.e., high efficiency levels), but also in a lower position in the region of high efficiency scores (i.e., low efficiency levels), thus it is ambiguous to conclude which sub-group has a more favourable density.

Next, we examine the confidence intervals of sub-group efficiencies. With the bootstrap approach, it is not statistically sufficient to conclude which sub-group is more efficient.

[21] Results for sub-group efficiencies based on hospital size and geographical location are provided in the Appendix.

However, using the central limit theorem approach, we see that the aggregate efficiency of the non-teaching sub-group is significantly lower (i.e., more efficient) than the aggregate efficiency of the teaching sub-group. Specifically, the confidence interval for the non-teaching sub-group, which ranges from 1.58 to 1.97 (i.e., efficiency levels from 0.51 to 0.63), is on the left of and does not overlap with the confidence interval for the teaching subgroup, which ranges from 2.04 to 2.35 (i.e., efficiency levels from 0.43 to 0.49). The result is consistent with those in Grosskopf et al. (2001) and Chowdhury and Zelenyuk (2016), who also compare the efficiency of teaching and non-teaching hospitals relative to the CRS frontier. Specifically, Grosskopf et al. (2001) benchmark teaching hospitals against the production frontier estimated using a non-teaching hospital sample and show that about 90% of the teaching hospitals could not efficiently "compete" with the "best practice" non-teaching hospitals. Chowdhury and Zelenyuk (2016) utilize the double bootstrap truncated regression to examine the impact of a set of explanatory variables including teaching status on hospital efficiency and find that teaching hospitals are, on average and ceteris paribus, less efficient than non-teaching hospitals.

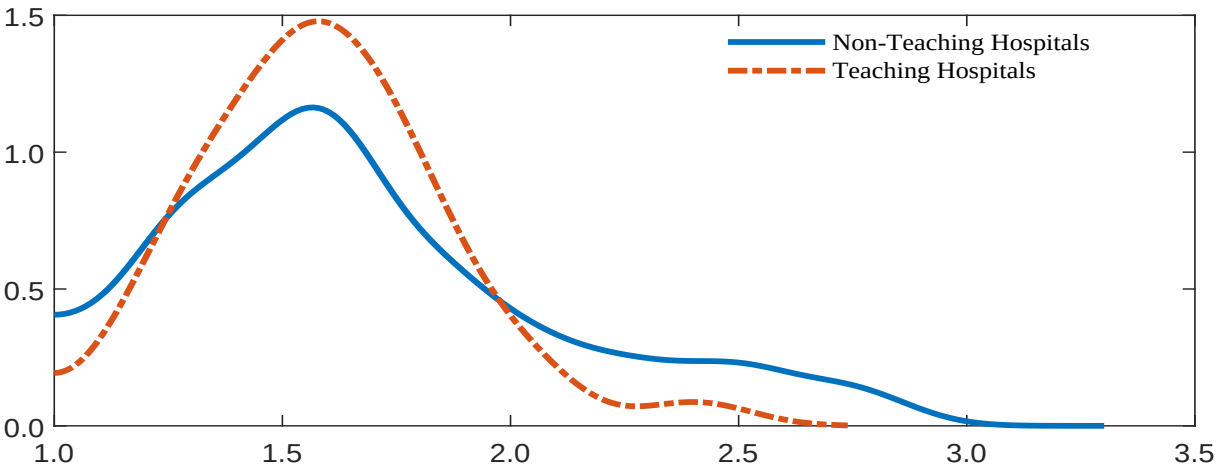

Figure 6: Density estimates of the estimated efficiency scores using the CRS-DEA estimators: non-teaching vs. teaching hospitals. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

Table 6: CRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Non-teaching and Teaching Hospitals

		Aggregate Efficiency		Simple Mean	
		Non-teaching	Teaching	Non-teaching	Teaching
Bootstrap	DEA estimate	1.53	1.61	1.66	1.59
	Bias corrected	1.79	2.02	1.96	1.96
	Est. Std	0.06	0.14	0.07	0.12
	LB of est. CI	1.65	1.60	1.81	1.63
	UB of est. CI	1.91	2.14	2.10	2.09
CLT	DEA estimate	1.53	1.61	1.66	1.59
	Bias corrected	1.83	2.19	1.96	2.10
	Est. Std	0.10	0.06	0.04	0.02
	Bias corrected Std	0.10	0.08	0.04	0.05
	LB est. CI	1.58	2.08	1.85	2.05
	UB of est. CI	1.96	2.30	1.99	2.15
	LB of est. CI-Improved	1.58	2.04	1.83	1.99
	UB of est. CI-Improved	1.97	2.35	2.01	2.22

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

4.3.2 Results for VRS-DEA

As in the case of the CRS-DEA estimator, we first estimate and compare the density of estimated efficiency scores of the each sub-group. Figure 7 shows that the teaching sub-group has a more favourable density compared to the non-teaching sub-group, i.e., its density is above in the area of low efficiency scores (i.e., high efficiency levels) and is below in the area of high efficiency scores (i.e., low efficiency levels).

Looking at confidence intervals, one can see that the results using the VRS-DEA estimator are opposite to those obtained using the CRS-DEA estimator. The teaching sub-group turns out to be significantly more efficient than the non-teaching sub-group regardless of the measures and approaches employed to compare them. For example, with the bootstrap approach, the confidence interval for the aggregate efficiency of the teaching sub-group, which ranges from 1.11 to 1.16 (i.e., efficiency levels from 0.86 to 0.90), is on the left of and does not overlap with the confidence interval for the aggregate efficiency of the non-teaching sub-group, which ranges from 1.48 to 1.65 (i.e., efficiency levels from 0.61 to 0.68). Similarly, the confidence interval estimated using the central limit theorem

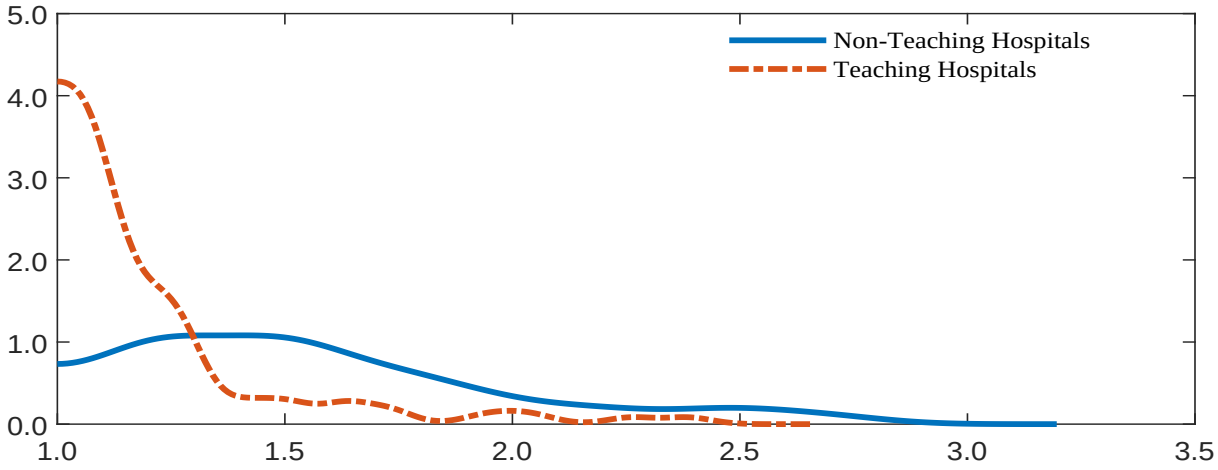

Figure 7: Density estimates of the estimated efficiency scores using VRS-DEA estimators: non-teaching vs. teaching hospitals. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

approach for the aggregate efficiency of the teaching sub-group, which ranges from 1.12 to 1.29 (i.e., efficiency levels from 0.77 to 0.89), is also on the left of and does not overlap with the confidence interval for the aggregate efficiency of the non-teaching sub-group, which ranges from 1.29 to 1.74 (i.e., efficiency levels from 0.57 to 0.77). Interestingly, the result is consistent with the finding in Nayar et al. (2013) – a study in which the VRS-DEA approach was employed. Specifically, Nayar et al. (2013) utilize the VRS input-oriented and slack-based additive DEA model with Tobit regression to examine the efficiency of acute care hospitals in the U.S and find that among other factors, the teaching status is positively related to hospital efficiency level.[22]

4.3.3 CRS vs. VRS

As teaching hospitals are mainly large hospitals, the differences in the conclusions about the relative efficiency of teaching and non-teaching subgroups when benchmarking to CRS frontier compared to VRS frontier can be explained by the scale inefficiency of the teaching subgroup.[23] Indeed, all except one observation in the teaching sub-group exhibit diseconomies of scale with the average scale inefficiency level of 0.27. Meanwhile, the average scale inefficiency level of the non-teaching subgroup is only 0.07. The result that teaching hospitals are less likely to be at optimal scale compared to non-teaching hospitals

[22]It is worth mentioning here that the use of Tobit regression in two-stage DEA context is not appropriate (see more discussion in Simar & Wilson, 2007).

[23]In our sample, 81% of teaching hospitals are large hospitals.

Table 7: VRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Non-teaching and Teaching Hospitals

		Aggregate Efficiency		Simple Mean	
		Non-teaching	Teaching	Non-teaching	Teaching
Bootstrap	DEA estimate	1.37	1.09	1.55	1.18
	Bias corrected	1.57	1.14	1.85	1.28
	Est. Std	0.04	0.01	0.06	0.04
	LB of est. CI	1.48	1.11	1.72	1.20
	UB of est. CI	1.65	1.16	1.95	1.34
CLT	DEA estimate	1.37	1.09	1.55	1.18
	Bias corrected	1.65	1.18	1.92	1.32
	Est. Std	0.11	0.04	0.05	0.03
	Bias corrected Std	0.11	0.04	0.07	0.04
	LB est. CI	1.30	1.13	1.69	1.29
	UB of est. CI	1.73	1.28	1.90	1.42
	LB of est. CI-Improved	1.29	1.12	1.65	1.28
	UB of est. CI-Improved	1.74	1.29	1.93	1.47

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

can also be found in Grosskopf et al. (2001).

It also appears to be the fact that the high efficiency level of teaching hospitals under the VRS assumption is because those hospitals are so large that there are not many (or even any) peers around them to be compared to reveal their true inefficiency. Absence of such peers also implies less (or no) competition, meaning less of the natural force from "the invisible hand" (noted by Adam Smith's), that much like gravity in physics, should pressure the decision making units (hospitals here) to be more efficient and more productive. If it is the case, then is it justified to deem them as efficient from a social point of view? It also raises natural questions for policy makers: Should the teaching hospitals be so large? And more generally, should the policy makers, instead of expanding already large public hospitals, invest into building new hospitals that are near the socially-optimal scale of operations? It is largely an open question for the judgement of experts and up for a healthy debate. Here, instead of directly answering the question, let us discuss an example to rationalise the importance of a socially-optimal scale for hospitals.

For simplicity, consider the single-input-single-output example that is easy to illustrate

pictorially. Suppose we observe six hospitals, namely A, B, C, D, E, and F with their input-output combinations being $(5;3)$, $(5;3)$, $(10;10)$, $(20;15)$, $(30;17)$, and $(40;17)$ respectively (Figure 8). All the six hospitals A, B, C, D, E, and F are on the boundary of the VRS technology, but only hospital B is on the CRS frontier–the frontier that identifies the best practice socially-optimal scale. One can see that society can benefit from utilizing resources more efficiently and provide relatively more services from relatively less resources. For example, it can have 4 more units of hospital outputs by merging hospital A and hospital B. More gains can be attained by splitting hospital D into two smaller hospitals of the same size as hospital C, thus producing 5 more units of hospital output without increasing the total number of inputs. Even many more gains can be attained when splitting hospital E or especially hospital F into, respectively, 3 or 4 smaller hospitals of the same size as hospital C, i.e, the society has, respectively, 13 or even 23 more units of hospital output with the same utilization of inputs. Therefore, from a social point of view, hospitals D, E, and F use too many resources to deliver what can be otherwise done by smaller size hospitals using much less resources.

Indeed, operating at the socially-optimal scale is of particular importance when the healthcare systems get sudden shocks, like pandemics, e.g., the one experienced by the world as this paper is being scribbled. The hospitals that operate at the socially-optimal scale are more flexible to expand their operations, while those that operate on the decreasing scale already, especially if they are far from the optimal scale like the hospital F in Figure 8, may have a hard time expanding their operations or will do so with even greater drainage or inefficient use of resources to deliver the necessary healthcare services to society.

Moreover, by becoming larger and larger, some organisations (including hospitals) may be turning into "too big to fail" entities that are not only experiencing the diseconomies of scale (as we see evidence here) and thus defeating the idea of their growth, but also suppressing the growth of others. Very large organizations may also tend to exhibit a weaker collegial culture of interpersonal professional relationships (Indik, 1963; Ingham, 1967). Indeed, and in general, when organizations become so big that an average employee is implicitly viewed like a little bolt in a big machine rather than an important member of a team, the productivity and efficiency of such organisations can be jeopardised leading to waste of resources for society.

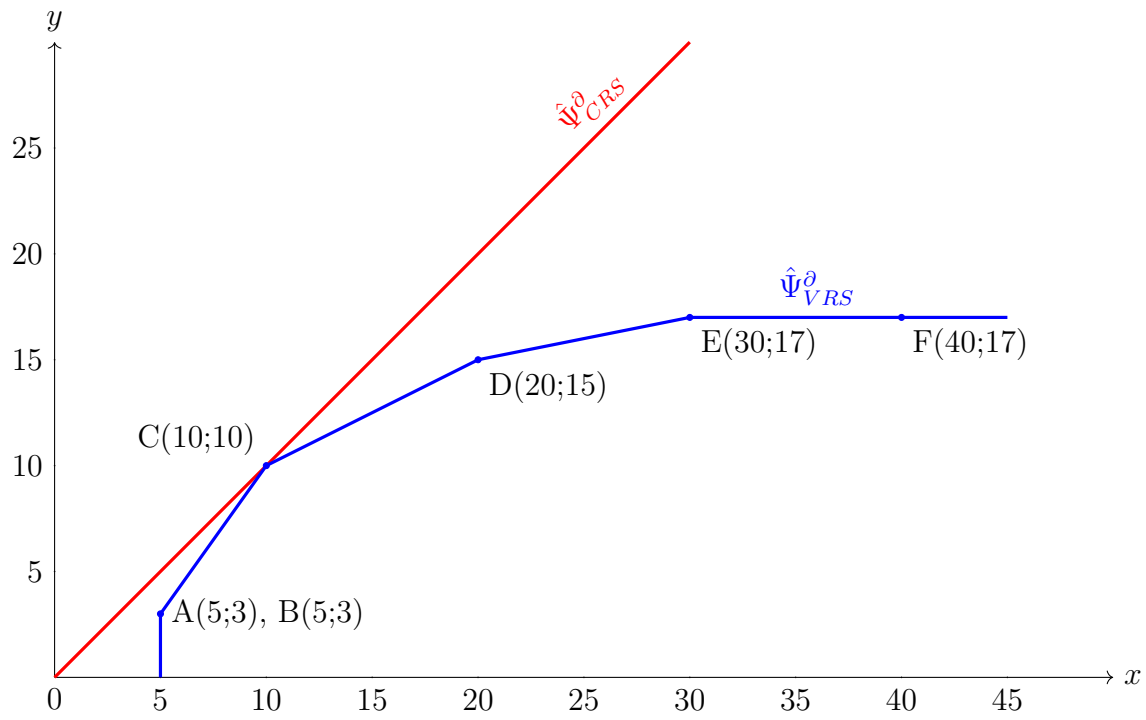

Figure 8: Socially-optimal scale illustration

5 Concluding Remarks

This paper explores the efficiency of different groups of hospitals in Queensland, Australia, focusing on teaching and non-teaching hospitals, by adapting the most recent developments on statistical analysis of aggregate efficiency. We focus on the two approaches: the bootstrap approach proposed by Simar and Zelenyuk (2007) and the central limits theorems recently developed by Simar and Zelenyuk (2018, 2020). To adapt these developments, we had to extend the central limit theorems to the context where there are several sub-groups in the population, which is a modest theoretical novelty of this paper. Moreover, this paper is the first real use of these methods (with some novel extensions) for an empirical study.

We found that the conclusions about the relative efficiency between teaching and non-teaching hospitals dramatically depend on the reference technology. Specifically, when benchmarking to the constant returns to scale frontier, teaching hospitals are significantly less efficient than non-teaching hospitals. However, teaching hospitals are significantly more efficient than non-teaching hospitals when benchmarking to the variable returns to scale frontier. The difference is largely explained by the diseconomies of scales of teaching hospitals.

It is worth mentioning here that the focus of this paper was on analysing unconditional

moments (simple and equally weighted) as well as unconditional densities, and thus a future direction of research would be to analyse conditional moments. For example, this can be done using the truncated regression with the bootstrap approach of Simar and Wilson (2007) or other alternatives.[24] Another fruitful direction for future research would be to use robust (order-α or order-m) frontiers, with or without conditioning on other variables.[25]

Acknowledgments

The authors acknowledge the support from their institution. We also acknowledge the financial support from the Australian Research Council (from the ARC Future Fellowship grant FT170100401). We thank David Du, Hong Ngoc Nguyen, Zhichao Wang and Evelyn Smart for their feedback from proofreading. We acknowledge and thank Queensland Health for providing part of the data that we used in this study. These individuals and organizations are not responsible for the views expressed in this paper.

[24] A work in progress using this approach is currently being done by Grosskopf et al. (2020).

[25] A work in progress using this approach is currently being done by Nguyen and Zelenyuk (2020).

Appendix

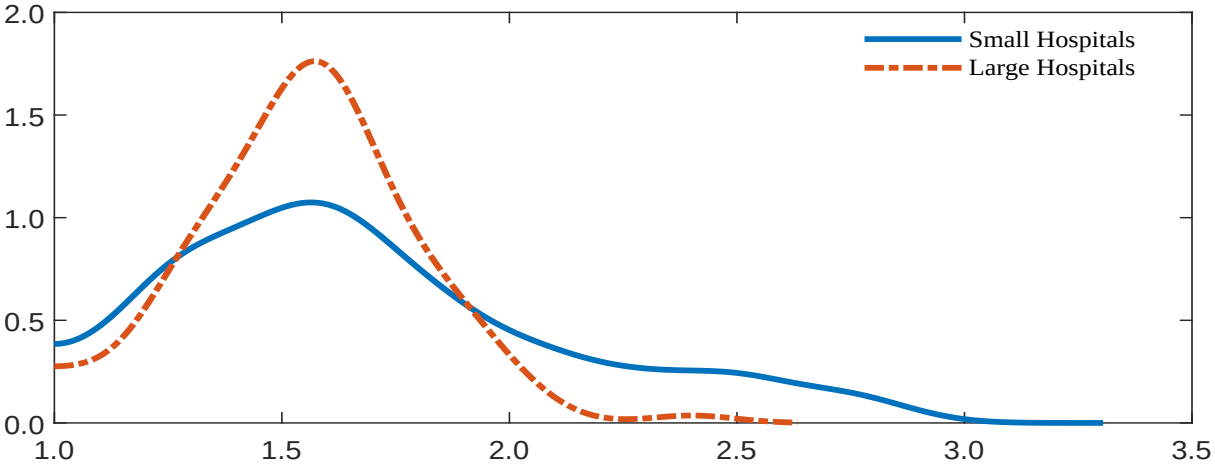

Figure 9: Density estimates of the estimated efficiency scores using the CRS-DEA estimators: small vs. large hospitals. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

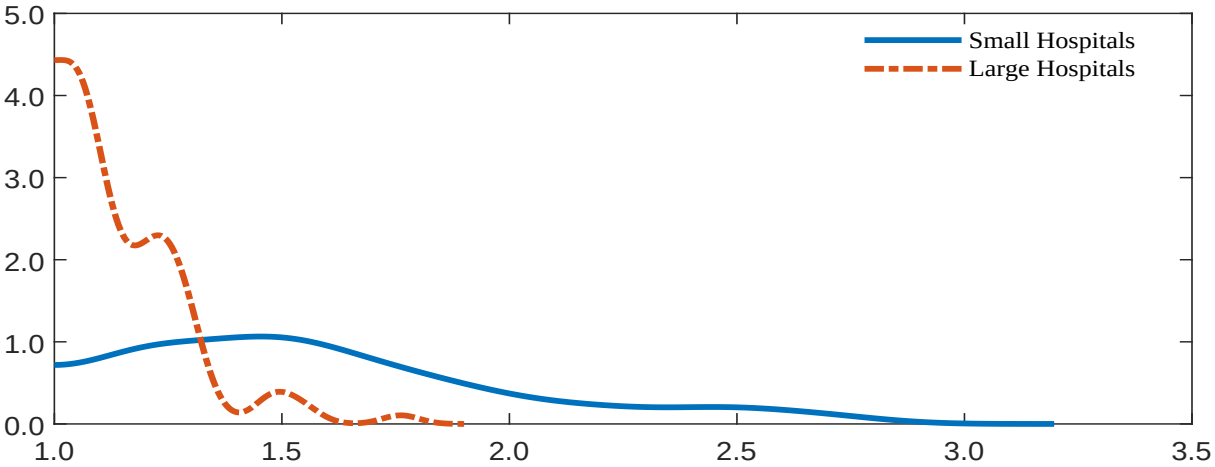

Figure 10: Density estimates of the estimated efficiency scores using the VRS-DEA estimators: small vs. large hospitals. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

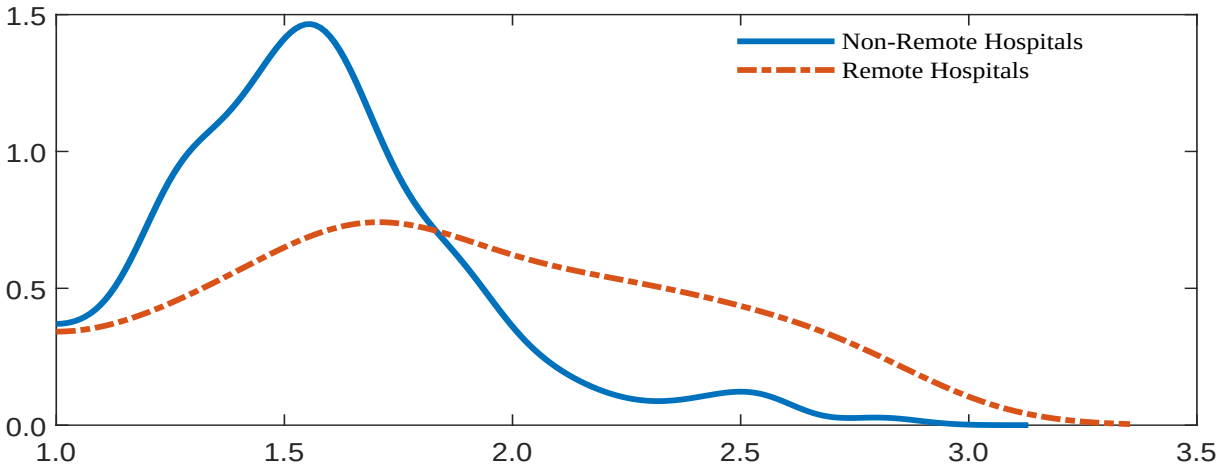

Figure 11: Density estimates of the estimated efficiency scores using the CRS-DEA estimators: non-remote vs. remote hospitals. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

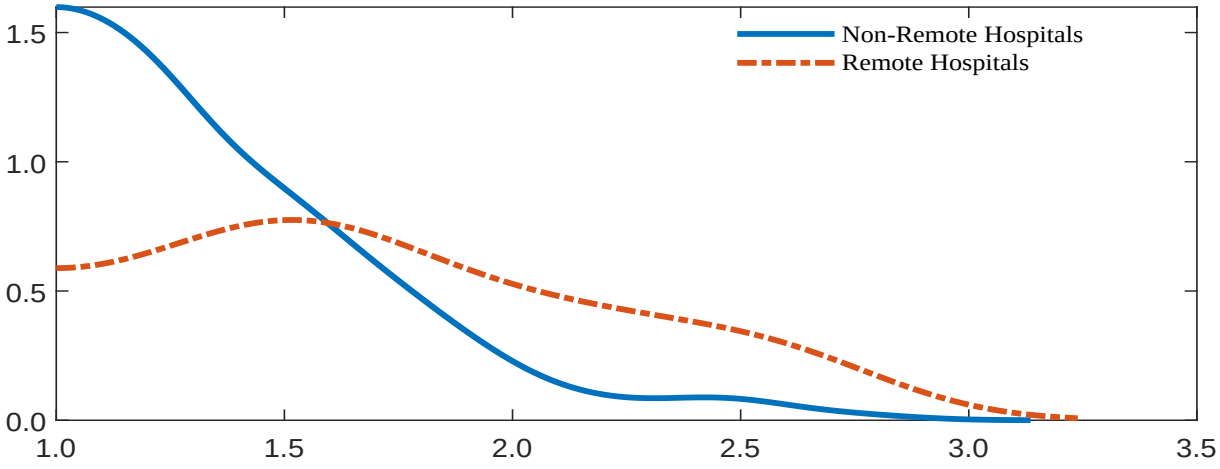

Figure 12: Density estimates of the estimated efficiency scores using the VRS-DEA estimators: non-remote vs. remote hospitals. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

Table 8: CRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Small and Large Hospitals

		Aggregate Efficiency		Simple Mean	
		Large	Small	Large	Small
Bootstrap	DEA estimate	1.60	1.56	1.55	1.68
	Bias corrected	1.98	1.81	1.90	1.97
	Est. Std	0.16	0.07	0.13	0.07
	LB of est. CI	1.58	1.68	1.58	1.82
	UB of est. CI	2.12	1.94	2.02	2.10
CLT	DEA estimate	1.60	1.56	1.55	1.68
	Bias corrected	2.16	1.81	2.06	1.94
	Est. Std	0.05	0.07	0.02	0.04
	Bias corrected Std	0.07	0.07	0.05	0.04
	LB est. CI	2.06	1.66	2.02	1.86
	UB of est. CI	2.27	1.93	2.10	2.01
	LB of est. CI-Improved	2.02	1.66	1.96	1.85
	UB of est. CI-Improved	2.31	1.94	2.16	2.01

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

Table 9: VRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Small and Large Hospitals

		Aggregate Efficiency		Simple Mean	
		Large	Small	Large	Small
Bootstrap	DEA estimate	1.09	1.46	1.13	1.57
	Bias corrected	1.15	1.69	1.20	1.87
	Est. Std	0.01	0.05	0.02	0.06
	LB of est. CI	1.12	1.57	1.16	1.75
	UB of est. CI	1.17	1.78	1.23	1.98
CLT	DEA estimate	1.09	1.46	1.13	1.57
	Bias corrected	1.19	1.75	1.25	1.94
	Est. Std	0.03	0.09	0.02	0.05
	Bias corrected Std	0.04	0.10	0.02	0.07
	LB est. CI	1.12	1.55	1.22	1.78
	UB of est. CI	1.26	1.91	1.29	1.99
	LB of est. CI-Improved	1.12	1.53	1.21	1.75
	UB of est. CI-Improved	1.26	1.93	1.30	2.03

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

Table 10: CRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Non-remote and Remote Hospitals

		Aggregate Efficiency		Simple Mean	
		Non-remote	Remote	Non-Remote	Remote
Bootstrap	DEA estimate	1.59	1.76	1.57	1.86
	Bias corrected	1.99	2.12	1.86	2.22
	Est. Std	0.14	0.11	0.07	0.11
	LB of est. CI	1.59	1.88	1.69	1.99
	UB of est. CI	2.12	2.33	1.98	2.41
CLT	DEA estimate	1.59	1.76	1.57	1.86
	Bias corrected	2.16	2.27	1.90	2.25
	Est. Std	0.11	0.17	0.03	0.05
	Bias corrected Std	0.12	0.18	0.04	0.06
	LB est. CI	1.94	1.93	1.84	2.16
	UB of est. CI	2.36	2.60	1.95	2.34
	LB of est. CI-Improved	1.92	1.92	1.82	2.14
	UB of est. CI-Improved	2.38	2.62	1.97	2.35

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

Table 11: VRS-DEA Estimates of Aggregate Efficiency and Simple Mean Efficiency and their 95% Confidence Intervals for Non-Remote and Remote Hospitals

		Aggregate Efficiency		Simple Mean	
		Non-remote	Remote	Non-remote	Remote
Bootstrap	DEA estimate	1.12	1.43	1.38	1.70
	Bias corrected	1.20	1.63	1.58	2.06
	Est. Std	0.01	0.09	0.05	0.09
	LB of est. CI	1.16	1.44	1.48	1.86
	UB of est. CI	1.22	1.77	1.66	2.21
CLT	DEA estimate	1.12	1.43	1.38	1.70
	Bias corrected	1.24	1.72	1.63	2.15
	Est. Std	0.09	0.12	0.05	0.06
	Bias corrected Std	0.09	0.13	0.05	0.09
	LB est. CI	1.04	1.43	1.53	2.02
	UB of est. CI	1.40	1.92	1.71	2.27
	LB of est. CI-Improved	1.04	1.42	1.51	1.97
	UB of est. CI-Improved	1.40	1.93	1.73	2.30

CLT: Central Limit Theorems, LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std: Standard Deviation, CI: Confidence Interval.

References

Australian Institute of Health and Welfare. (2015). *Australian hospital peer groups* (tech. rep. No. 66). Australian Institute of Health and Welfare. Canberra, ACT.

Ayanian, J. Z., & Weissman, J. S. (2002). Teaching hospitals and quality of care: A review of the literature. *The Milbank Quarterly*, *80*(3), 569–593.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, *30*(9), 1078–1092.

Berta, P., Callea, G., Martini, G., & Vittadini, G. (2010). The effects of upcoding, cream skimming and readmissions on the Italian hospitals efficiency: A population-based investigation. *Economic Modelling*, *27*(4), 812–821.

Besstremyannaya, G. (2013). The impact of Japanese hospital financing reform on hospital efficiency: A difference-in-difference approach. *The Japanese Economic Review*, *64*(3), 337–362.

Bickel, P. J., & Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 967–985.

Biørn, E., Hagen, T. P., Iversen, T., & Magnussen, J. (2003). The effect of Activity-Based Financing on hospital efficiency: A Panel Data Analysis of DEA efficiency scores 1992–2000. *Health Care Management Science*, *6*(4), 271–283.

Burgess, J. F., & Wilson, P. W. (1996). Hospital ownership and technical inefficiency. *Management Science*, *42*(1), 110–123.

Burgess, J. F., & Wilson, P. W. (1998). Variation in inefficiency among US hospitals. *INFOR: Information Systems and Operational Research*, *36*(3), 84–102.

Cameron, J. M. (1985). The indirect costs of graduate medical education. *New England Journal of Medicine*, *312*(19), 1233–1238.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*(6), 429–444.

Chowdhury, H., & Zelenyuk, V. (2016). Performance of hospital services in Ontario: DEA with truncated regression approach. *Omega*, *63*, 111–122.

Chua, C. L., Palangkaraya, A., & Yong, J. (2011). Hospital competition, technical efficiency and quality. *Economic Record*, *87*(277), 252–268.

Clement, J. P., Valdmanis, V. G., Bazzoli, G. J., Zhao, M., & Chukmaitov, A. (2008). Is more better? an analysis of hospital outcomes and efficiency with a DEA model of output congestion. *Health Care Management Science*, *11*(1), 67–77.

Daraio, C., & Simar, L. (2007). Economies of scale, scope and experience in the italian motor-vehicle sector. In C. Daraio & L. Simar (Eds.), *Advanced robust and nonparametric methods in efficiency analysis: Methodology and applications* (pp. 135–165). Springer Science & Business Media.

Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency in post offices. In M. G. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurements* (pp. 243–267). Amsterdam, North-Holland.

Färe, R., Grosskopf, S., & Logan, J. (1983). The relative efficiency of illinois electric utilities. *Resources and Energy*, *5*(4), 349–367.

Färe, R., He, X., Li, S., & Zelenyuk, V. (2019). A unifying framework for farrell profit efficiency measurement. *Operations Research*, *67*(1), 183–197.

Färe, R., & Primont, D. (1995). *Multi-output production and duality: Theory and applications* (R. Färe & D. Primont, Eds.). New York: Kluwer Academic Publishers.

Färe, R., & Zelenyuk, V. (2003). On aggregate Farrell efficiencies. *European Journal of Operational Research*, *146*(3), 615–620.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, *120*(3), 253–290.

Ferrier, G. D., & Trivitt, J. S. (2013). Incorporating quality into the measurement of hospital efficiency: A double DEA approach. *Journal of Productivity Analysis*, *40*(3), 337–355.

Grosskopf, S., Margaritis, D., & Valdmanis, V. (2001). Comparing teaching and non-teaching hospitals: A frontier approach (teaching vs. non-teaching hospitals). *Health Care Management Science*, *4*(2), 83–90.

Grosskopf, S., Nguyen, B. H., Yong, J., & Zelenyuk, V. (2020). Healthcare structural reform and the performance of public hospitals: The case of Queensland, Australia [In Progress].

Hao, S. S., & Pegels, C. C. (1994). Evaluating relative efficiencies of Veterans Affairs medical centers using Data Envelopment, ratio, and multiple regression analysis. *Journal of Medical Systems*, *18*(2), 55–67.

Harris, J., Ozgen, H., & Ozcan, Y. (2000). Do mergers enhance the performance of hospital efficiency? *The Journal of the Operational Research Society*, *51*(7), 801–811.

Hofmarcher, M. M., Paterson, I., & Riedel, M. (2002). Measuring hospital efficiency in Austria – a DEA approach. *Health Care Management Science*, *5*(1), 7–14.

Independent Hospital Pricing Authority. (2013). *National efficient price determination 2013-2014* (tech. rep.). Independent Hospital Pricing Authority.

Indik, B. P. (1963). Some effects of organization size on member attitudes and behavior. *Human Relations*, *16*(4), 369–384.

Ingham, G. K. (1967). Organizational size, orientation to work and industrial behaviour. *Sociology*, *1*(3), 239–258.

Jensen, G. A., & Morrisey, M. A. (1986). The role of physicians in hospital production. *The Review of Economics and Statistics*, 432–442.

Kneip, A., Park, B. U., & Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, *14*, 783–793.

Kneip, A., Simar, L., & Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, *24*(6), 1663–1697.

Kneip, A., Simar, L., & Wilson, P. W. (2015). When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores. *Econometric Theory*, *31*(2), 394–422.

Lehner, L. A., & Burgess, J. F. (1995). Teaching and hospital production: The use of regression estimates. *Health Economics*, *4*(2), 113–125.

Magnussen, J. (1996). Efficiency measurement and the operationalization of hospital production. *Health services research*, *31*(1), 21–37.

Nayar, P., & Ozcan, Y. A. (2008). Data Envelopment Analysis comparison of hospital efficiency and quality. *Journal of Medical Systems*, *32*(3), 193–199.

Nayar, P., Ozcan, Y. A., Yu, F., & Nguyen, A. T. (2013). Benchmarking urban acute care hospitals: Efficiency and quality perspectives. *Health Care Management Review*, *38*(2), 137–145.

Nguyen, B. H., & Zelenyuk, V. (2020). Robust efficiency analysis of public hospitals in Queensland, Australia [In Progress].

O'Neill, L., Rauner, M., Heidenberger, K., & Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences*, *42*(3), 158–189.

Park, B. U., Jeong, S.-O., & Simar, L. (2010). Asymptotic distribution of conical-hull estimators of directional edges. *The Annals of Statistics*, *38*(3), 1320–1340.

Park, B. U., Simar, L., & Weiner, C. (2000). FDH efficiency scores from a stochastic point of view. *Econometric Theory*, *16*, 855–877.

Politis, D. N., Romano, J. P., & Wolf, M. (2001). On the asymptotic theory of subsampling. *Statistica Sinica*, 1105–1124.

Productivity Commission. (2010). *Public and private hospital: Multivariate analysis* (tech. rep. Supplement to Research Report). Productivity Commission. Canberra, ACT66.

Rich, E. C., Gifford, G., Luxenberg, M., & Dowd, B. (1990). The relationship of house staff experience to the cost and quality of inpatient care. *Journal of the American Medical Association*, *263*(7), 953–957.

Shahian, D. M., Nordberg, P., Meyer, G. S., Blanchfield, B. B., Mort, E. A., Torchiana, D. F., & Normand, S.-L. T. (2012). Contemporary performance of us teaching and nonteaching hospitals. *Academic Medicine*, *87*(6), 701–708.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 683–690.

Shephard, R. W. (1953). *Cost and production functions*. Princeton University Press.

Shephard, R. W. (1970). *Theory of cost and production functions*. Princeton University Press.

Sickles, R., & Zelenyuk, V. (2019). *Measurement of productivity and efficiency*. Cambridge University Press.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London ; New York, Chapman; Hall.

Simar, L., & Wilson, P. W. (2007). Estimation and inference in two-stage, semi-parametric models of production processes. *Journal of Econometrics*, *136*(1), 31–64.

Simar, L., & Wilson, P. W. (2011). Inference by the m out of n bootstrap in nonparametric frontier models. *Journal of Productivity Analysis*, *36*(1), 33–53.

Simar, L., & Zelenyuk, V. (2006). On testing equality of distributions of technical efficiency scores. *Econometric Reviews*, *25*(4), 497–522.

Simar, L., & Zelenyuk, V. (2007). Statistical inference for aggregates of Farrell-type efficiencies. *Journal of Applied Econometrics*, *22*(7), 1367–1394.

Simar, L., & Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research*, *66*(1), 137–149.

Simar, L., & Zelenyuk, V. (2020). Improving finite sample approximation by central limit theorems for estimates from data envelopment analysis. *European Journal of Operational Research*.

Worthington, A. C. (2004). Frontier efficiency measurement in health care: A review of empirical techniques and selected applications. *Medical Care Research and Review*, *61*(2), 135–170.

Zelenyuk, V. (2020). Aggregation of inputs and outputs prior to data envelopment analysis under big data. *European Journal of Operational Research*, *282*(1), 172–187.

Zelenyuk, V., & Zheka, V. (2006). Corporate governance and firm's efficiency: The case of a transitional country, Ukraine. *Journal of Productivity Analysis*, *25*(1–2), 143–157.

Zuckerman, S., Hadley, J., & Lezzoni, L. (1994). Measuring hospital efficiency with frontier cost functions. *Journal of Health Economics*, *13*(3), 255–280.