XIX

Volume XIX, Number 1, May 2016

Journal of Applied Economics

Richard Gearhart

The robustness of cross-country healthcare rankings among homogeneous OECD countries

UCEMA

Edited by the Universidad del CEMA
Print ISSN 1514-0326
Online ISSN 1667-6726

Journal of Applied Economics. Vol XIX, No. 1 (May 2016), 113-144

THE ROBUSTNESS OF CROSS-COUNTRY HEALTHCARE RANKINGS AMONG HOMOGENEOUS OECD COUNTRIES

Richard Gearhart*
California State University, Bakersfield

Submitted June 2014; accepted June 2015

This paper re-examines analyses of cross-country healthcare efficiency using modern, non-parametric estimators and Malmquist indices to determine productivity changes over the panel. This paper finds that cross-country heterogeneity leads to different efficiency rankings than previously thought, and that the hyperbolic order-α estimator leads to more robust efficiency scores when looking across different output measures, only looking at the more homogeneous OECD countries. It finds that the United States, if excluding the percent of healthcare expenditures that are publicly financed, is one of the more inefficient healthcare delivery systems in the world. What are commonly thought of as well-run healthcare systems (Austria and France) are either inefficient themselves or have variation in their efficiency rankings, showcasing difficulties in using other countries' healthcare systems as models for reform. It also finds that there has been productivity regression in all countries except the United States. These highlight the difficulties in cross-country efficiency comparisons.

JEL classification codes: C14, I11, I12, I18
Key words: United States healthcare, production efficiency, order- α, DEA, OECD

I. Introduction

The debate over the relative efficiency of the U.S. healthcare system is still not settled. According to the National Research Council and Institute of Medicine (2013), even though America is one of the wealthiest countries in the world, it is far from the healthiest. The report notes that life expectancy for men and women in America is near the worst among developed countries, and the prevalence of

---

* Richard Gearhart: Assistant Professor of Economics, School of Business and Public Administration, Department of Economics, California State University, Bakersfield, 20 BDC, California State University, Bakersfield; 9001 Stockdale Highway, Bakersfield, CA 93311, USA; email rgearhart1@csub.edu. I would like to thank Paul W. Wilson for helpful comments throughout the process. I would also like to thank two anonymous referees and Mariana Conte Grand (a co-editor) for many valuable comments. All errors are my own.

certain diseases, such as heart disease and diabetes, is much higher (National Research Council and Institute of Medicine 2013). These rankings can be misleading, as these health outcomes are determined by factors outside of health care. For instance, a comparatively high rate of fatal car accidents and murders in the U.S. bias the life expectancy number when compared to other developed countries (Pipes 2013). Pipes (2013) notes how the report from the Institute of Medicine states that the percentage of pre-term births is exceptionally high when compared to developed European countries.[1] The aggressiveness of U.S. doctors in trying to save pre-term babies biases certain measures of health outcomes and can lead to observed differences in health trends across countries (Pipes 2013).[2]

The debate about the relative merits of U.S. health care compared to the rest of the developed world is still unsettled. This leads to the need for reliable efficiency rankings to drive health care policy. Without the ability to derive consistent efficiency rankings across estimators or datasets, policy may lead to more inefficient systems or unanticipated results, which can lead to different health care delivery system outcomes from the ex-ante objectives for initiating change. This can also lead to method-searching, where policymakers will choose the set of results that best support their stated goals.

This paper is motivated as (1) a re-examination of the original World Health Organization (WHO) dataset using newly available non-parametric estimation methods in the hyperbolic direction, rather than the input and output direction, that were not available in previous studies (Evans et al. 2000; Tandon et al. 2000; Hollingsworth and Wildman 2003; Greene 2004; Afonso and St. Aubyn 2005), to allow for comparisons of efficiency rankings, (2) an empirical analysis, using non-parametric estimation methods, to Organization for Economic Cooperation and Development (OECD) countries only over a broader time span, and (3) a robustness check of minimal alterations in outputs used, with the same non-parametric estimators and input mix, on efficiency rankings across countries.[3]

[1] The percent of pre-term births in America is 65% higher than that of Great Britain, and about double the rates in Finland and Greece.

[2] This occurs even though the U.S. neo-natal mortality rate has dropped from almost 95% in the 1960s to about 5% today.

[3] The input direction means finding what reduction in inputs is feasible given an output level being produced. The output direction means finding what expansion in outputs is feasible given an input level being fixed. The hyperbolic direction means finding the combinations of input reduction and output expansion that are still feasible.

My models are different from previous studies in a variety of dimensions: (1) compared to previous studies that use non-parametric estimators (Hollingsworth and Wildman 2003; Afonso and St. Aubyn 2005), I am able to ignore the arbitrary choice between the input-orientation and the output-orientation, and work in a hyperbolic orientation, using either the data envelopment analysis (DEA) estimator (where previous estimates had no choice but to utilize either the input-orientation or the output-orientation) as well as use a newer non-parametric estimator, the order-α estimator, that, unlike the DEA estimator, does not suffer from the presence of outliers nor the curse of dimensionality; (2) compared to previous studies that use parametric estimators (Evans et al. 2000; Tandon et al. 2000; Greene 2004), the newer non-parametric order-α does not require distributional assumptions on its function form, and it seamlessly incorporates multiple outputs in the model; (3) the use of the hyperbolic-orientation Malmquist index, where previous studies (Hollingsworth and Wildman 2003) used the output-orientation Malmquist index, which can suffer from missing efficiency values if technology improves; and (4) the inclusion of alternative specifications on the inputs and outputs used, that aggregate from a wide variety of papers several potential (and reasonable) specifications to determine the efficiency of a healthcare delivery system.

The World Health Report 2000 (WHO, 2000) defines three intrinsic goals of the health system of a nation to be: (1) to improve health, (2) to be responsive to the legitimate demands of the population, and (3) to ensure that no one is at risk of serious financial losses because of ill health. Evans et al. (2000) and Tandon et al. (2000) used this to present a rankings-based comparison of the productive efficiency of the health care system of 191 countries. The preferred methodologies by Evans et al. (2000) and Tandon et al. (2000) were a fixed effects linear regression model, where the country-specific constants embody technical efficiency, which remain constant over time. Greene (2004) proposed several different alternative methodologies, including a stochastic frontier model and a random effects model. Based on his estimates, Greene (2004) noted the large effects that cross-country heterogeneity played in the rankings, and a large disparity in efficiency estimate rankings among OECD and non-OECD countries.

Hollingsworth and Wildman (2003) likewise extended the original WHO study, using both a parametric stochastic frontier estimator and a DEA estimator to account for heterogeneity in the data. Afonso and St. Aubyn (2005) estimated efficiency of OECD countries using DEA and free disposal hull (FDH) estimators.

They focus on both input-oriented and output-oriented efficiency measures, and find that the United States is not nearly as inefficient as earlier studies suggested.

Richardson et al. (2003) criticized the output measures selected by the World Health Organization (WHO), and the weights assigned to the output measures. The first output measure, Disability Adjusted Life Expectancy (DALE), is a measure of healthy life expectancy, based on morbidity (Mathers et al. 2000; Richardson et al. 2003). The second output measure, Composite Measure of Health Care Delivery (COMP), is composed of multiple components. The five components for the COMP output criterion are: (1) Maximizing population health (DALE), (2) Reducing inequalities in population health, (3) Maximizing health system responsiveness, (4) Reducing inequalities in health system responsiveness, and (5) Financing health care equitably. When using DEA, each of these measures was given a weight, with (1), (2), and (5) receiving weights of 0.25, while (3) and (4) received weights of 0.125. Unfortunately, DEA estimators suffer from their own weighting problem, where the optimization problem utilized by the efficiency calculation weights inputs and outputs so that the decision-making unit (here, the country) maximizes efficiency. This means that inefficient inputs and outputs may be given a weight of 0 during the optimization, potentially biasing the results (Allen et al. 1997; Charnes et al. 1978; Coelli et al. 2005).

For the output measures themselves, Williams (2001) criticized the DALE and COMP measures for relying heavily on speculative assumptions made throughout the survey and pointed out that many of the health care statistics were interpolated (for the U.S., the only statistic that was not interpolated was child mortality). Williams (2001) also suggests that using DALE or COMP introduces tremendous bias into the estimates. Garber and Skinner (2008) note that cross-country comparisons of the efficiency estimates suffer from structural differences in the countries themselves (i.e., obesity rates due to food consumption patterns).

Greene (2004) noted that the OECD countries as a whole are significantly different from non-OECD countries, and suggests looking at OECD countries solely. Berger and Messer (2002) constructed an analysis similar to that of the WHO, but looked solely at OECD countries. They focused on a much narrower base of health care inputs and outputs. The overall mortality rate was used as the proxy for health outcomes in a given year, and aggregate health care expenditures per capita were used as an input. They included dummy variables to control for country and year fixed effects, but their work solely focused on parametric regression analysis. Berger and Messer (2002) argue that one of the most basic ways to alter health care delivery systems is to change public funding of health care expenditures. It is important to have more reliable and valid efficiency results so that policy implications of a reduced (or expanded) public influence in the health sector can be more accurately analyzed. Kim and Kang (2014) note that the selection of appropriate inputs and outputs to measure cross-country healthcare efficiency is considered difficult in healthcare research. In a snapshot of 11 studies on efficiency measures of healthcare systems from 2003 to 2011 (Table 1 in their paper), they find that no two studies had the same input- and output-specifications.

These concerns lead to doubts about the robustness of cross-country healthcare comparisons, in practice and for reform, even among the more homogeneous OECD countries. I show the robustness of cross-country healthcare efficiency rankings with a newer, non-parametric estimator across different output measures; the hyperbolic order- estimator.[4] Even though the hyperbolic order-α estimator eliminates some of the technical concerns found using the DEA estimators, the efficiency rankings must still be interpreted with caution, however, as cross-country heterogeneity still plays a large role in the efficiency estimates if utilizing a set of countries with significant differences (such as looking at all countries). Compared to previous parametric estimates of healthcare efficiency by Evans et al. (2000), Tandon et al. (2000), and Greene (2004), hyperbolic order-α estimator efficiency rankings are robust to changes in the output mixture when looking at OECD countries only, a homogeneous grouping. For instance, changing from the COMP output measure to the DALE output measure leads the U.S. to increase its rankings from 30th (dead last) to 29th. Switching to an output mixture that includes infant survival rate and a measure of life expectancy leads the U.S. to be ranked 27th (dead last).

In contrast, Greene (2004) found that switching from the COMP to the DALE output measure meant the U.S. improved its rankings, going from 25th to 13th in the OECD. I conclude that the more technically advanced, newly developed non-parametric estimators provide more robust rankings across the output measures commonly considered for cross-country healthcare comparisons since 2000. When the countries are more homogeneous, as suggested by Greene (2004)

[4] There are other new non-parametric estimators that have been developed, such as the order-m estimator. See Cazals et al. (2002) for details. Order-m estimators were not used in the current paper because although the theoretical properties have been described in Wilson (2011), the ability to empirically apply these estimators is still in their infancy.

to look at OECD countries only, the robustness of the efficiency rankings are confirmed. The Pearson correlation coefficient between models with different output measures ranges from 0.657 to 0.97, using OECD countries only. Including more heterogeneous, non-OECD, countries in the sample reduces the robustness of the results. Preliminary evidence shows that the inclusion of a single additional input variable, the percent of health care expenditures that are publicly financed, rather than per capita healthcare expenditures, results in the U.S. being in the top quarter of OECD countries in terms of healthcare efficiency.[5]

II. Data

Data for this study include data used by Greene (2004), as well as data from the OECD and World Bank websites, and from the Barro-Lee "Educational Attainment in the World from 1950-2010" website.[6] The data include observations from 191 countries from the years 1993-1997 and 30 OECD countries from 1997-2006. There are some initial problems with the data, and assumptions must be made to make the analysis tractable.

I use one assumption for the data that has been used in previous works.[7] For educational attainment, Barro and Lee only observe data in five year intervals for the average years of schooling found from the Barro-Lee dataset on educational attainment. Thus, there are data from 1995, 2000, 2005, and 2010. The annual observations are linearly interpolated values of how educational attainment changes. The unobserved values of other inputs and the infant survival rate are interpolated using the same methodology as used for average educational attainment.[8] Table 1 shows the 5 different models used.

[5] Note that the inclusion of this additional input variable introduces difficulties in the analysis. Dyson et al. (2001) discuss this point.

[6] The dataset is located on the internet at http://www.barrolee.com/

[7] The Statistical Annex to the World Health Report 2000 notes that educational attainment values are imputed using a variety of methods.

[8] This assumption is based upon the realization that, if we were to delete all observations which had missing values (a non-interpolated data set), the efficiency estimates of both the interpolated data and non-interpolated data would lie around the 45 degree line, which means that we can use either interpolated data or non-interpolated data. To provide more data points, we use interpolated data. The figure to support this assumption is available upon request from the author.

Table 1. Explanations and comparisons of models used

	Model	Outputs	Inputs	Countries	Years Covered	Estimator Used
GREENE DATASET	Model 1	COMP	Per Capita Health Care Expenditures, Average Educational Attainment	OECD only	1993-1997	Non-parametric hyperbolic order-α
	Model 2	DALE	Per Capita Health Care Expenditures, Average Educational Attainment	OECD only	1993-1997	Non-parametric hyperbolic order-α
	Model 3	COMP	Per Capita Health Care Expenditures, Average Educational Attainment	All 191	1993-1997	Non-parametric hyperbolic order-α
	Model 4	DALE	Per Capita Health Care Expenditures, Average Educational Attainment	All 191	1993-1997	Non-parametric hyperbolic order-α
PREVIOUS STUDIES	Evans et al. (2000)	COMP	Per Capita Health Care Expenditures, Average Educational Attainment	All 191	1993-1997	Parametric fixed effects
	Tandon et al. (2000)	DALE	Per Capita Health Care Expenditures, Average Educational Attainment	All 191	1993-1997	Parametric fixed effects
	Greene (2004)	COMP	Per Capita Health Care Expenditures, Average Educational Attainment	All 191	1993-1997	Parametric stochastic frontier analysis (SFA)
	Greene (2004)	DALE	Per Capita Health Care Expenditures, Average Educational Attainment	All 191	1993-1997	Parametric stochastic frontier analysis (SFA)
OECD DATASET	Model 5	DALY, ISR	Per Capita Health Care Expenditures, Average Educational Attainment	OECD only	1997-2006	Non-parametric hyperbolic order-α
	Model 5 Alternative	DALY, ISR	Per Capita Health Care Expenditures, Average Educational Attainment, % Healthcare Expenditures Publicly Financed	OECD only	1997-2006	Non-parametric hyperbolic order-α

Notes: The countries used for estimation are either all 191 countries or OECD countries only. Models 1 through 4 are the models directly comparable to previous studies in that they use the same years and the same input-output combinations. Models 1 through 4 will be referred to as the Greene dataset. Model 5 differs in that it uses different years (1997 through 2006) with slightly altered output measures and input measures. Model 5 and Model 5 Alternative will be referred to as the OECD dataset.

The first 4 models are a direct re-estimation of the original Evans et al. (2000) and Tandon et al. (2000) analysis, reporting efficiency scores both relative to OECD countries only (as Greene 2004 suggests) and relative to all 191 countries (as Evans et al. 2000, Tandon et al. 2000, and Greene 2004 did). Thus, the first 4 models are based on earlier studies (Evans et al. 2000, Tandon et al. 2000, Greene 2004), using the same input (health care expenditures per capita, educational attainment) and output (DALE, COMP) measures, but using a different non-parametric estimator, the hyperbolic order-α frontier estimator unavailable to Evans et al. (2000), Tandon et al. (2000), and Greene (2004). I also provide explanations of the models used in previous studies.

In Model 5, I extend the analysis, using an expanded dataset with more years (1997 to 2006), using different output measures, using only OECD countries in the dataset. The output measures I use are the infant survival rate (infants who survive per 1,000 live births), ISR, and the fraction of years that are spent without disease, disability, or premature death, 1 - DALY, where DALY represents Disability Adjusted Life Years. DALY is a measure of both morbidity and mortality, a measure of overall disease burden, expressed as the fraction of years lost (per 100 year life span) due to ill health, disability, or premature death. This measure is an alternative measure of average life expectancy, conditional on the disease or disability burden of a country (whereas average life expectancy is an unconditional measure).[9] Proposed inputs are the same as in the original analysis; health care expenditures per capita and educational attainment levels.

The inclusion of an alternative input variable, the percent of healthcare expenditures that are publicly financed instead of per capita health care expenditures, has been considered in previous papers. Greene (2004) utilizes the percent of healthcare expenditures that are publicly financed as an additional environmental variable in his stochastic frontier analysis, noting that it is particularly important for health policy, though there is a tremendous amount of variation in this measure

[9] This is analogous to looking at the survival rates of patients across hospitals, but not controlling for the types of patients that enter the hospital for treatment. Thus, a hospital that is a cancer center will have a lower survival rate than a hospital that specializes in GI disorders (due to the nature of the patients that come to the hospital). Thus, what may seem to be differences in efficiency (due to differences in the survival rates) may simply be due to heterogeneity among the patients, which any output measure (such as the survival rate) needs to be conditioned on. This measure has its own limitations discussed earlier.

across OECD countries. Dyson et al. (2000) note that the inclusion of this input in the analysis creates potential biases in the efficiency estimates. This has not stopped the use of this as an input variable in previous studies, however. Adam et al. (2011) use the percent of healthcare expenditures that are publicly financed as an input variable in their DEA estimation. Kim and Kang (2014) do likewise. The fact that this input variable may artificially alter the efficiency rankings obtained because of bias leads to it not being included for most of the analysis in the paper.

One problem with the data is that, with at most 30 observations per year, and using two output and two input measures, this will lead to perhaps substantial estimation error in the estimates. The convergence rate of the DEA estimator will be slower than the convergence rate of parametric estimators (Kneip et al., 1998). From Kneip et al. (1998), the convergence rate of the DEA estimator is $n^{-\frac{2}{(p+q+1)}}$, where p is the number of inputs and q is the number of outputs, while the convergence rate of parametric estimators is $n^{-\frac{1}{2}}$.[10] This problem is mitigated using the hyperbolic order-α estimator. A second problem with the data is measurement error. Much of the data obtained from the OECD is obtained from analyses conducted by individual governments. Methodological differences exist and can lead to faulty conclusions when comparing raw numbers.[11]

I estimate efficiency for each country using the hyperbolic order-α estimator.[12] I then compare estimated efficiency (for any country) between any 2 years, decomposing changes in productivity using a Malmquist index. I cannot simply use the difference between efficiency scores in adjacent years as a measure of technological growth, because the change in efficiency scores could be due to a number of causes. It could be that there is the same input-output combination but different technologies; the same technology but different input-output combinations; or some combinations of different technologies and different input-output combinations.

[10] The convergence rate of my DEA estimator is $n^{-\frac{2}{5}}$. With 30 observations per year, this is analogous to having $30^{-\frac{2}{5}} = m^{-\frac{1}{2}}$observations, or m = 15.19 observations per year with a parametric estimator with normal root-n convergence.

[11] The United States counts an infant exhibiting any sign of life as alive, no matter the month of gestation or size. France, Ireland, and the Netherlands (among others) do not report live births of babies under 500 grams and/or 22 weeks of gestation (MacDorman and Mathews 2009).

[12] Results for the DEA estimators are not reported but available upon request from the author.

III. Estimators

This paper uses hyperbolic non-parametric estimators for a variety of reasons. Parametric estimators, like those used by Evans et al. (2000), require potentially untenable specification assumptions or have other serious drawbacks. Stochastic frontier analysis (SFA), a parametric estimator based on the ideas of Aigner et al. (1977) and Meeusen and von den Broeck (1977), involves estimation of a function with a composite error term. This composite error term consists of both inefficiency and noise components, making empirical distinction between the two quite hard. Likewise, use of SFA estimators require assumptions about the distribution of this composite error term, often by a half-normal or truncated normal distribution. Another drawback to traditional parametric estimators is the difficulty in incorporating multiple outputs in an analysis.

Non-parametric estimators are often used by researchers because they do not require an a priori specification of the functional relationship that is being estimated. Similarly, because of the lack of distributional assumption, incorporating multiple outputs is seamless. However, certain non-parametric estimators, such as the DEA estimator used by Afonso and St. Aubyn (2005), suffer from well-known problems that make validity and inference a problem. The problems include the DEA estimator having less than root-n convergence due to the curse of dimensionality, where the number of observations required to obtain meaningful estimates increases with the number of production inputs and outputs used in the estimation, and the estimator being sensitive to outliers (Kneip et al. 1998).

Alternatives to DEA estimators, such as the order-α estimator (which involves estimating a partial frontier lying "close" to the true production frontier), have been developed which alleviate many of these problems.[13] Unlike the DEA estimator, the order-α estimator is robust to outliers. The order-α estimator is a partial frontier estimator, and it allows some observations to lie above the estimated partial frontier, limiting the impacts of extreme values (or outliers) on efficiency scores (Simar and Wilson 2008). The order-α estimator also addresses the curse of dimensionality found in DEA estimators; by design, the order-α estimator achieves the classical, parametric root-n rate of convergence, even though it is

[13] Aragon et al. (2005) and Daouia and Simar (2007) estimate input- and output-oriented conditional quantiles of order-α.

a fully non-parametric estimator (Wheelock and Wilson 2009). The hyperbolic order-α estimator thus provides the distributional flexibility of non-parametric estimators while simultaneously providing traditional statistical features found in parametric estimators.[14] Other estimators, such as the order-m estimator, have also been developed to alleviate problems found in the DEA estimator, though they are not used in the current paper (Cazals et al. 2002).

Unfortunately, having to choose between an input-orientation and an output-orientation leads to an issue surrounding the order-α estimator as well as the DEA estimator. Afonso and St. Aubyn (2005) note that input-orientation and output-orientation estimates will only be the same under constant returns to scale. Afonso and St. Aubyn (2005) chose to provide estimates in both the input-orientation and the output-orientation, while Hollingsworth and Wildman (2003) provided estimates in the output-orientation only. Often, as noted in Wheelock and Wilson (2009), the choice between input- or output-orientation is often arbitrary. Wheelock and Wilson (2008) offer a way out the choice between the input-orientation and the output-orientation. They describe an unconditional hyperbolic order-α quantile estimator that shares the advantages of the estimators described in Aragon et al. (2005) and Daouia and Simar (2007), but which avoids the third problem of choosing the orientation of the estimator. Since I am outside the context of a regression framework, the choice of direction function (input, output, or hyperbolic) does not have behavioral implications as it does in regression analysis; I thus use the hyperbolic distance function. This allows for input contraction at a given output level, output expansion at a given input level, or a combination of input contraction and output expansion.

Due to this, I utilize the hyperbolic order-α estimator, which is a partial frontier estimator. The order-α estimator was developed a potential solution to the known problems in other non-parametric estimators described above, and where $\alpha \in (0,1]$ corresponds to the level of an appropriate non-standard conditional quantile frontier. The choice of α is continuous on the interval (0,1]. Wheelock and Wilson (2008) define the hyperbolic order-α estimator as

$$\gamma_\alpha(x,y) = sup\{\gamma > 0 | H(\gamma^{-1}x, \gamma y) > (1-\alpha)\} \tag{1}$$

[14] These statistical features are a lack of sensitivity to outliers as well as the root- n convergence rate.

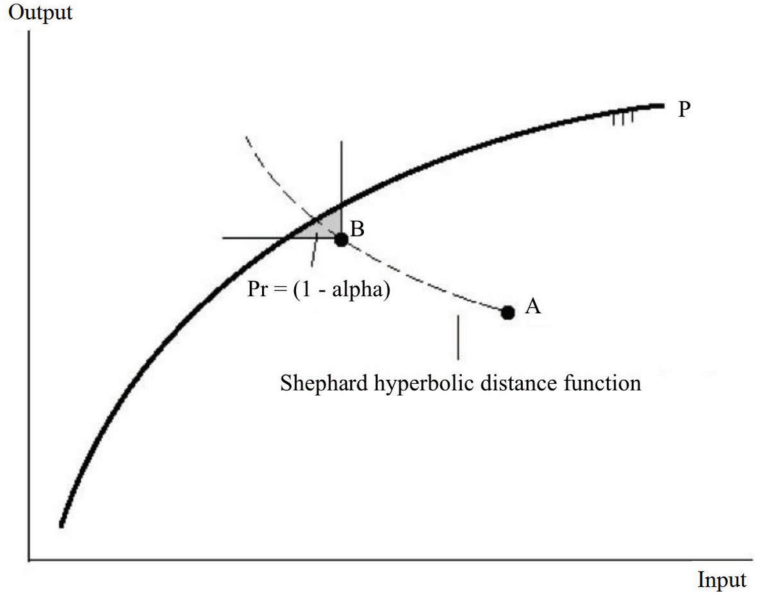
using the Shephard (1970) metric, where $H(x,y) = \Pr(X \leq x, Y \geq y)$, which represents the probability that a decision-making unit operating at (x,y) is dominated by another decision-making unit producing more output with the same level of inputs; producing the same level of output with less inputs; or producing more outputs using less inputs). I estimate $H(x,y)$by $\widehat{H}(x,y) = \sum_{i=1}^{n}\left(\frac{I(X_i \leq x, Y_i \geq y)}{n}\right)$, where $I(\cdot)$ represents the indicator function. I estimate γ_α by

$$\widehat{\gamma_\alpha}(x,y) = sup\{\gamma > 0 \big| \widehat{H}(\gamma^{-1}x, \gamma y) > (1-\alpha)\}. \tag{2}$$

Wheelock and Wilson (2008) establish the consistency of the hyperbolic order-α estimator.

If $\gamma_\alpha(x,y) = 1$, the point is said to lie on the hyperbolic order-α quantile and is dominated by a decision-making unit with a probability of $(1-\alpha)$ (Simar and Wilson, 2008). Figure 1 illustrates an order-α estimator.

Figure 1. Graphical representation of hyperbolic order-α estimator



A decision-making unit at point A could contract inputs, expand outputs, and move to point B, and only have a $((1-\alpha)100)\%$ chance of being dominated. This is represented by the shaded area in Figure 1. I also create a 95% confidence interval for each hyperbolic order-α efficiency point estimate using a smooth bootstrap (Wheelock and Wilson 2008). In addition, another useful feature of the order-α estimator is that the estimator has an asymptotic normal distribution (Wheelock and Wilson 2008).

I use the Malmquist index decomposition proposed in Wheelock and Wilson (1999), which helps to determine changes in productivity, efficiency, scale, and technology over time. Malmquist indices are estimated using the DEA Shephard (1970) hyperbolic, output, or input distance functions. I use the hyperbolic distance function because of problems caused when calculating a Malmquist index in the input or output direction. The problem is that the Malmquist indices are calculated using DEA estimates and, because of the way that they are constructed, the estimates may not have values. This happens when, due to a technological shock that shifts the production frontier, the observed data point may lie outside of the estimated frontier. When this happens, the DEA estimate will be indeterminate and cannot be calculated. This problem occurs in the input or output directions, and is eliminated through the use of the Shephard (1970) hyperbolic distance function.

The Malmquist index decomposition proposed by Wheelock and Wilson (1999), between any two years t_1 and t_2, where $t_1 < t_2$, is found to be

$$M(t_1,t_2) = \left(\frac{D_i^{t_2|t_2}}{D_i^{t_1|t_1}}\right)\cdot\left[\frac{D_{i,CRS}^{t_2|t_2}}{D_i^{t_2|t_2}}\cdot\frac{D_i^{t_1|t_1}}{D_{i,CRS}^{t_1|t_1}}\right]\cdot\left[\frac{D_i^{t_2|t_1}}{D_i^{t_2|t_2}}\cdot\frac{D_i^{t_1|t_1}}{D_i^{t_1|t_2}}\right]^{\frac{1}{2}}\cdot\left[\frac{D_{i,CRS}^{t_2|t_1}}{D_i^{t_2|t_1}}\frac{D_i^{t_2|t_2}}{D_{i,CRS}^{t_2|t_2}}\cdot\frac{D_{i,CRS}^{t_1|t_1}}{D_i^{t_1|t_1}}\frac{D_i^{t_1|t_2}}{D_{i,CRS}^{t_1|t_2}}\right]^{\frac{1}{2}}\cdot \quad (3)$$

I define $D_i^{t_j|t_k}\ \forall j,k$ to be the hyperbolic distance function from the ith country's position in the input/output space at time t_j to the boundary of the production set at time t_k, while defining $D_{i,CRS}^{t_j|t_k}\ \forall j,k$ to be the hyperbolic distance function from the ith country's position in the input/output space at time to the convex cone of the production set at time t_k.

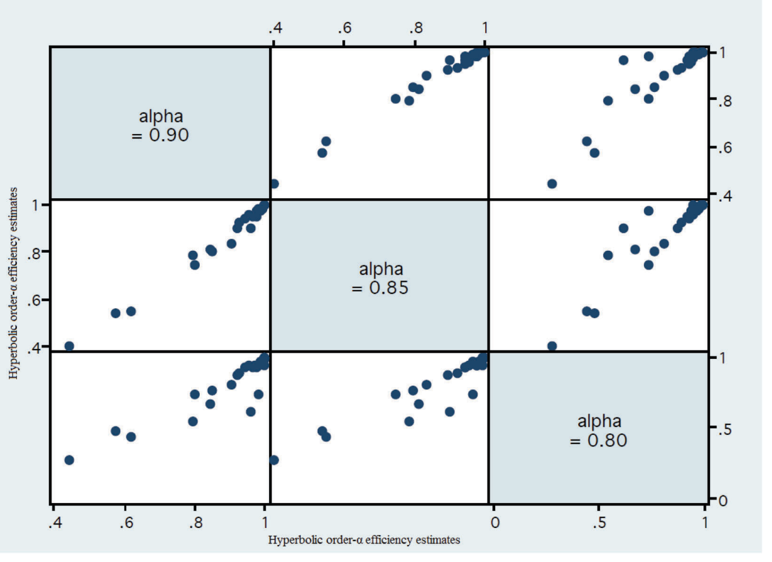
Thus, the total value $M = M_i(t_1,t_2)$ represents the change in productivity for the ith country between any two years; the value $PE = \left(\frac{D_i^{t_2|t_2}}{D_i^{t_1|t_1}}\right)$ represents the change in pure efficiency for the ith country between any two years (captures how the production input-output combination changes); the value $S = \left[\frac{D_{i,CRS}^{t_2|t_2}}{D_i^{t_2|t_2}}\cdot\frac{D_i^{t_1|t_1}}{D_{i,CRS}^{t_1|t_1}}\right]$

represents the change in scale for the ith country between any two years; the value $PT = \left[\frac{D_i^{t_2|t_1}}{D_i^{t_2|t_2}} \cdot \frac{D_i^{t_1|t_1}}{D_i^{t_1|t_2}}\right]^{\frac{1}{2}}$ represents the change in pure technology for the ith country between any two years (captures how the production frontier changes); and the value $ST = \left[\frac{D_{i,CRS}^{t_2|t_1}}{D_i^{t_2|t_1}} \frac{D_i^{t_2|t_2}}{D_{i,CRS}^{t_2|t_2}} \cdot \frac{D_{i,CRS}^{t_1|t_1}}{D_i^{t_1|t_1}} \frac{D_i^{t_1|t_2}}{D_{i,CRS}^{t_1|t_2}}\right]^{\frac{1}{2}}$ represents the change in the scale of technology for the ith country between any two years (captures how technology becomes flatter, i.e., the production frontier exhibits more CRS, or becomes more curved, i.e., the production frontier exhibits more VRS). I define M to be the Malmquist index (productivity), PE to be pure efficiency, S to be scale, PT to be pure technology, and ST to be scale technology. In the hyperbolic direction, when the measures (M, PE, S, PT, ST) take the range of values $< (=, >) 1$ this indicates improvement (no change, regression) in the measurement. One problem with the inference of the Malmquist index is that, because it uses a kernel density estimator when bootstrapping values, it suffers from the curse of dimensionality (Simar and Wilson 1999).

IV. Results

The interpretation of the hyperbolic order-α efficiency scores is as follows: an efficiency score of δ means that the country uses 100 δ-percent of the inputs and produces 100 δ-percent of the output as a country that lies on the production frontier. Scores of less than 1 indicates a country is more efficient than others, while scores of greater than 1 indicates the country is more inefficient than others. The lower the efficiency score, the more efficient a country is. In empirical applications, for the hyperbolic order-α estimator, a value for α must be chosen to construct the estimated frontier. Efficiency estimates seem to be robust to the choice of α, as shown in Figure 2.

Figure 2. Different values of α for hyperbolic order-α efficiency estimator, model 1, 1997



Note: Figure 2 plots the hyperbolic order-α efficiency estimates for 3 values of α — 0.8, 0.85, and 0.9 — against each other using Model 1. Each panel compares estimates for a pair of values for α.

Figure 2 plots the hyperbolic α-quantile efficiency estimates for three values of α — 0.8, 0.85, and 0.9 — against each other; each panel compares estimates of a pair of values for α. There is a similar ranking of countries across different values of α due to the fact that most points fall on (or near) a straight line. I use a value of α =0.90. That these results seems to be robust with respect to the choice of α is seen in other studies (Daouia and Simar 2007; Wheelock and Wilson 2008, 2009).

I compute order-α efficiency estimates for the countries from the dataset used in Evans et al. (2000), Tandon et al. (2000), and Greene (2004), for each year 1993 to 1997 using the Frontier Efficiency Analysis with R software package in R. I present the results for 1997 in Table 2.[15]

[15] Efficiency estimates for all other years available upon request from the author.

Table 2. Hyperbolic quantile efficiency estimates, Greene dataset, 1997, α = 0.90

	OECD Countries only				All 191 countries			
	Model 1 (COMP)		Model 2 (DALE)		Model 3 (COMP)		Model 4 (DALE)	
Country	Estimate	Rank	Estimate	Rank	Estimate	Rank	Estimate	Rank
Australia	0.994	18	0.9994	18	0.9322	7	0.9332	5
Austria	0.959	11	0.9688	13	0.9868	26	0.9839	24
Belgium	0.9957	20	1	19	0.9593	16	0.9586	13
Canada	0.9989	26	1.006	26	0.9659	20	0.9623	14
Czech Republic	0.9513	9	0.9565	10	0.9173	4	0.9467	6
Denmark	1.0084	29	1.0455	30	0.9917	27	1.0006	28
Finland	0.9972	25	1	19	0.9352	9	0.9474	7
France	0.9543	10	0.9486	9	0.9847	24	0.9638	16
Germany	1.0036	28	1.0297	28	0.9945	28	1.0007	29
Greece	0.8974	7	0.884	5	0.9099	2	0.9084	1
Hungary	0.8904	6	0.8904	6	0.9341	8	0.9855	25
Iceland	0.9946	19	1	19	0.9644	19	0.969	20
Ireland	0.9722	13	0.9937	16	0.9285	6	0.9557	12
Italy	0.9068	8	0.9068	7	0.9631	18	0.9539	10
Japan	0.9776	14	0.9607	11	0.9409	12	0.9294	4
Luxembourg	0.9959	22	1.0191	27	0.9837	23	0.9901	26
Mexico	0.791	2	0.791	2	0.9598	17	0.9784	22
Netherlands	0.9964	23	1	19	0.9796	21	0.9666	18
New Zealand	1.0013	27	1.0057	25	0.9397	10	0.9636	15
Norway	0.9787	15	0.9716	15	0.9426	13	0.9491	8
Poland	0.8754	4	0.9375	8	0.9067	1	0.9555	11
Portugal	0.8389	3	0.8389	3	0.9408	11	0.9497	9
Slovakia	0.9593	12	0.9635	12	0.9481	14	0.967	19
South Korea	0.9873	17	1	19	0.951	15	1	27
Spain	0.8774	5	0.8774	4	0.9263	5	0.9155	2
Sweden	0.9958	21	0.9959	17	0.9807	22	0.9648	17
Switzerland	0.9965	24	1.003	24	0.985	25	0.9724	21
Turkey	0.4505	1	0.4505	1	1.0143	30	0.9812	23
UK	0.983	16	0.9712	14	0.9163	3	0.9282	3
US	1.0103	30	1.0387	29	0.9977	29	1.0062	30

Notes: The first 4 columns reflect the efficiency estimates and rankings for the Greene dataset, Models 1 and 2, using OECD countries only to estimate the production frontier. The last 4 columns reflect the efficiency estimates and rankings for the Greene dataset, Models 3 and 4, using all 191 countries to estimate the production frontier, where the efficiency rankings are derived relative to OECD countries only. The interpretation of the efficiency estimates is as follows: an efficiency score of means that the country uses percent of the inputs and produces percent of the output as a country that lies on the production frontier. This means that scores less than 1 indicate a country is more efficient than others, while a score of more than 1 means the country is more inefficient than others.

The efficiency rankings differ on which output measure I use and if the data are a more heterogeneous sample (use 191 countries instead of OECD countries only). My rankings diverge from those in Evans et al. (2000), Tandon et al. (2000), and Greene (2004). This shows that careful selection of the input-output bundles and estimators used to measure the frontier is imperative. Utilizing the same framework as previous studies, with the same input-output combinations, but a different non-parametric estimator, I find that the U.S. continues to be highly inefficient. In fact, the rankings obtained show that the U.S. is one of the most inefficient countries in the world, worse than previous rankings obtained in the original study (Evans et al., 2000; Tandon et al., 2000). My estimates indicate that the U.S. is, at best, ranked 29th. Using the point estimates for the U.S. for Model 1, in 1997 the U.S. uses 101-percent of the inputs and produces $\left(\frac{100}{1.0103}\right) = 98.98$-percent of the output compared to a country located on the $\alpha = 0.90$ quantile frontier along the hyperbolic path from the United States.

Table 3 shows a comparison of the efficiency rankings from my estimates, and the previous estimates found in Evans et al. (2000), Tandon et al. (2000), and Greene (2004).

Table 3. Comparison of efficiency estimates from 1997 across studies

	COMP estimates				DALE estimates			
Country	Model 1	Model 3	Greene	Tandon et al.	Model 2	Model 4	Greene	Evans et al.
Australia	18	7	11	20	18	5	9	19
Austria	11	26	10	4	13	24	15	7
Belgium	20	16	15	15	19	13	12	14
Canada	26	20	13	18	26	14	8	18
Czech Republic	9	4	22	24	10	6	25	26
Denmark	29	27	18	21	30	28	24	23
Finland	25	9	16	19	19	7	22	21
France	10	24	9	1	9	16	4	2
Germany	28	28	19	17	28	29	21	20
Greece	7	2	1	8	5	1	2	5
Hungary	6	8	27	29	6	25	30	29
Iceland	19	19	14	9	19	20	20	13
Ireland	13	6	17	13	16	12	19	16
Italy	8	18	7	2	7	10	11	1
Japan	14	12	3	5	11	4	5	4
Luxembourg	22	23	21	10	27	26	1	15
Mexico	2	17	29	27	2	22	18	22
Netherlands	23	21	5	11	19	18	16	9
New Zealand	27	10	20	6	25	15	6	8
Norway	15	13	6	23	15	8	17	25
Poland	4	1	24	25	8	11	23	28
Portugal	3	11	23	7	3	9	26	6
Slovakia	12	14	26	28	12	19	14	27
South Korea	17	15	28	26	19	27	27	30
Spain	5	5	2	3	4	2	29	3
Sweden	21	22	4	16	17	17	3	10
Switzerland	24	25	12	14	24	21	10	12
Turkey	1	30	30	30	1	23	28	17
UK	16	3	8	12	14	3	7	11
US	30	29	25	22	29	30	13	24

Notes: The study by Tandon et al. (2000) was a follow-up to the original study by Evans et al. (2000), using the same estimation procedure but choosing to use, as an output measure, COMP instead of DALE.

This provides evidence of the considerable variation in parametric efficiency estimates found in earlier studies when using a more heterogeneous sample of countries. Moving from the COMP output measure in Greene (2004) to the DALE output measure in Greene (2004) leads to a change in the U.S. ranking from 25th (out of 30 countries) to 13th (out of 30 countries). My more restricted sample of OECD countries only leads to a decrease in U.S. efficiency rankings from 30th out of 30 countries (COMP output measure) to 29th out of 30 countries (DALE output measure). Calculating Pearson correlation coefficients shows that, even though efficiency rankings are highly correlated even with a more heterogeneous sample, they are more highly correlated looking at OECD countries only. This provides evidence that, using the hyperbolic order-α estimator on a more homogeneous country set leads to highly robust efficiency rankings across a variety of output measures. For example, the correlation coefficient between Models 1 and 2 is 0.97 when looking at OECD countries only. This compares to the coefficient between Models 3 and 4, which is 0.77, when looking at all 191 countries. There is even less correlation when looking at the estimates from Greene (2004); the correlation coefficient between his model with COMP and his model with DALE is 0.50. This again highlights the fact that newer non-parametric estimators provide highly robust results when looking at a homogeneous sample. Even a wholesale change in which output measures are being used, from COMP or DALE to infant survival rate and DALY, leads to a relatively high Pearson correlation coefficient of almost 0.70.

As Table 4 shows, the inclusion of an alternative input variable, the percent of healthcare expenditures that are publicly financed rather than per capita healthcare expenditures, alters the efficiency ranking for the United States considerably.[16]

The U.S. is now ranked, at worst, 4th among OECD countries, and is even the country with the most efficient healthcare delivery system under certain specifications (if all 191 countries are used with the COMP output measure). Both of these results highlight the extreme caution that must be used in even interpreting the hyperbolic order-α estimator.

[16] As noted previously, the inclusion of this input variable leads to potential biases in the estimates.

Table 4. Hyperbolic quantile efficiency estimates, Greene dataset, 1997, α = 0.90, with percent of healthcare expenditures publicly financed as alternative input

	Estimates OECD Countries only		Estimates All 191 Countries	
	Model 1 (COMP) with Alternative Input	Model 2 (DALE) with Alternative Input	Model 3 (COMP) with Alternative Input	Model 4 (DALE) with Alternative Input
Country	Rank	Rank	Rank	Rank
Australia	17	15	17	18
Austria	7	9	8	9
Belgium	25	22	21	17
Canada	16	13	16	20
Czech Republic	29	29	28	28
Denmark	27	27	25	27
Finland	18	18	24	24
France	6	5	9	7
Germany	19	21	20	22
Greece	10	7	10	10
Hungary	28	30	30	30
Iceland	15	17	22	21
Ireland	14	19	23	16
Italy	3	2	6	6
Japan	20	16	11	11
Luxembourg	26	25	19	23
Mexico	4	10	3	3
Netherlands	12	12	14	14
New Zealand	21	23	26	25
Norway	13	14	13	15
Poland	24	26	27	26
Portugal	5	4	5	5
Slovakia	2	8	2	2
South Korea	30	28	29	29
Spain	9	6	7	8
Sweden	22	20	15	13
Switzerland	11	11	12	12
Turkey	8	1	1	1
UK	23	24	18	19
US	1	3	4	4

Notes: This table re-examines Models 1 through 4 and the efficiency rankings found in Table 2 with the inclusion of an alternative input, the percent of healthcare expenditures that are publicly financed, rather than per capita healthcare expenditures. The inclusion of this alternative input has been found in other works, including Berger and Messer (2002) and Kim and Kang (2014).

In Table 5, I provide the point estimates and corresponding 95% confidence intervals for the models that use the DALE output measure, estimating efficiency with OECD countries only, and with all 191 countries.[17]

[17] Point estimates and the corresponding 95% confidence intervals for Models 1 and 3 are available upon request from the author.

Table 5. Hyperbolic quantile efficiency and CI estimates, Greene dataset, 1997, α = 0.90

	OECD Countries Only Model 2 (DALE output measure)			All 191 Countries Model 4 (DALE output measure)		
Country	Estimate	Lower CI	Upper CI	Estimate	Lower CI	Upper CI
Australia	0.9994	0.9412	1.1052	0.9332	0.6862	1.3349
Austria	0.9688	0.8953	1.0688	0.9839	0.7186	1.3492
Belgium	1	0.9421	1.1095	0.9586	0.7443	1.3411
Canada	1.006	0.9424	1.1113	0.9623	0.7341	1.3611
Czech Republic	0.9565	0.8642	1.0327	0.9467	0.7116	1.3619
Denmark	1.0455	1.0225	1.1958	1.0006	0.8298	1.4556
Finland	1	0.9553	1.1314	0.9474	0.6855	1.3448
France	0.9486	0.8721	1.0430	0.9638	0.6863	1.3281
Germany	1.0297	0.9891	1.1630	1.0007	0.7948	1.4413
Greece	0.884	0.7209	0.9002	0.9084	0.6030	1.2731
Hungary	0.8904	0.7982	0.9676	0.9855	0.7516	1.4658
Iceland	1	0.9454	1.1075	0.969	0.7498	1.3854
Ireland	0.9937	0.9225	1.0897	0.9557	0.7120	1.3645
Italy	0.9068	0.7885	0.9624	0.9539	0.6510	1.2619
Japan	0.9607	0.8636	1.0361	0.9294	0.6730	1.3047
Luxembourg	1.0191	0.9709	1.1396	0.9901	0.8012	1.4149
Netherlands	1	0.9344	1.1079	0.9666	0.7247	1.3595
New Zealand	1.0057	0.9488	1.1234	0.9636	0.7339	1.3662
Norway	0.9716	0.9004	1.0715	0.9491	0.7246	1.3265
Portugal	0.8389	0.6072	0.7740	0.9497	0.6493	1.2914
Slovakia	0.9635	0.8864	1.0533	0.967	0.7591	1.3915
South Korea	1	0.8522	1.0194	1	0.8246	1.4480
Spain	0.8774	0.7580	0.9342	0.9155	0.6279	1.2734
Sweden	0.9959	0.9201	1.0932	0.9648	0.7317	1.3606
Switzerland	1.003	0.9301	1.1039	0.9724	0.7276	1.3763
UK	0.9712	0.9010	1.0686	0.9282	0.6714	1.3538
US	1.0387	0.8976	1.0636	1.0062	0.7108	1.3863

Notes: The first 3 columns reflect the efficiency estimates and 95% confidence interval (CI) bounds for the Greene dataset for the DALE output measure only (Model 2), using OECD countries to estimate the production frontier. The last 3 columns reflect the efficiency estimates and 95% confidence interval bounds for the Greene dataset, for the DALE output measure only (Model 4), using all 191 countries to estimate the production frontier. The interpretation of the efficiency estimates is as follows: an efficiency score of means that the country uses percent of the inputs and produces percent of the output as a country that lies on the production frontier. This means that scores less than 1 indicate a country is more efficient than others, while a score of more than 1 means the country is more inefficient than others.

The 95% confidence interval for the efficiency estimates increases when we increase the sample size to include all 191 countries (Model 4) and not just the 30 OECD countries only (Model 2). This provides evidence for the notion that countries not in the OECD are quite heterogeneous from those in the OECD, and supports the idea that an analysis of cross-country healthcare efficiency must use data from OECD countries only. This supports previous analysis, as Greene (2004) and Hollingsworth and Wildman (2003) suggest looking at the OECD countries only, which seems to provide a more accurate depiction by looking at a more homogeneous sample. Note, however, that this is still a second-best solution; even reducing the number of countries to OECD only, it is unlikely that the efficiency rankings are significantly closer to the truth, and within-country analysis may be a better option.

In Table 6, I provide the rankings only for the extension of the original analysis, by using different output measures from the years 1997 to 2006 (Model 5).[18]

Table 6. Hyperbolic quantile efficiency estimates, OECD data, 1997, $\alpha = 0.90$

Country	Model 5 Rank	Model 5 Alternative Rank
Australia	21	13
Austria	18	14
Belgium	23	20
Canada	23	21
Czech Republic	9	25
Denmark	26	22
Finland	14	12
France	19	18
Germany	20	19
Greece	7	2
Hungary	5	27
Iceland	12	9
Ireland	22	23
Italy	6	6
Japan	10	8
Luxembourg	16	16
Netherlands	13	11
New Zealand	15	24
Norway	23	15
Poland	1	4
Portugal	1	4
Slovakia	2	26
South Korea	3	1
Spain	4	7
Sweden	11	10
Switzerland	17	5
UK	8	17
US	27	3

Notes: These estimates represent efficiency rankings for the OECD dataset (Model 5), using as output measures infant survival rate (ISR) and disability adjusted life years (DALY). Model 5 utilizes as inputs average educational attainment and per capita healthcare expenditures. Model 5 Alternative adds, as a third input, the percent of healthcare expenditures that are publicly financed.

[18] To be able to compare results to other models where the years of estimation are 1993 to 1997, I report results for 1997 from Model 5 and Model 5 Alternative only. Other years are available upon request from the author.

The outputs have been altered, to alleviate some of the concerns leveled by other authors (Richardson et al. 2003; Williams 2001). With the inclusion, as an input, of per capita healthcare expenditures, rather than the percent of health care expenditures that are publicly financed, the U.S. continues to be the most inefficient healthcare delivery system in the world, ranking 27th out of 27 OECD countries. However, as seen in the third column of Table 6, including the percent of healthcare expenditures that are publicly financed as an input along with per capita healthcare expenditures drastically alters the efficiency ranking for the U.S.; in 1997, the U.S. would have ranked 3rd out of 27 OECD countries. We see that the more plausible models, where the percent of healthcare expenditures is not included as an input in the production process, leads to the U.S. being one of the most inefficient producers of healthcare among the OECD and all 191 countries, as is shown in Table 7. This supports the common notion that American healthcare has significant issues that need to be addressed.

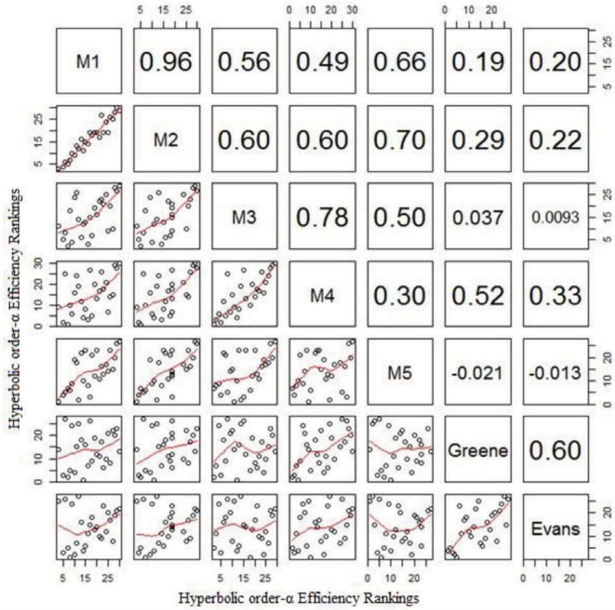
Table 7. Hyperbolic quantile efficiency ranks, U.S. only, α = 0.90

Year	Model 1	Model 2	Model 3 (OECD)	Model 3 (All)	Model 4 (OECD)	Model 4 (All)	Model 5
	Rank	Rank	Rank	Rank	Rank	Rank	Rank
1993	29/30	29/30	-	-	-	-	-
1994	28/30	29/30	-	-	-	-	-
1995	29/30	29/30	-	-	-	-	-
1996	30/30	29/30	-	-	-	-	-
1997	30/30	29/30	29/30	141/191	30/30	148/191	27/27
1998	-	-	-	-	-	-	27/27
1999	-	-	-	-	-	-	28/28
2000	-	-	-	-	-	-	27/27
2001	-	-	-	-	-	-	28/28
2002	-	-	-	-	-	-	28/28
2003	-	-	-	-	-	-	28/28
2004	-	-	-	-	-	-	27/27
2005	-	-	-	-	-	-	24/24

Notes: These are the efficiency estimate rankings of the U.S. over time. The number means that the U.S. is ranked ath out of b countries with observed efficiency rankings. All efficiency estimate rankings are relative to the other OECD countries (where observations are available), except in Models 3 and 4. These are the efficiency estimate rankings, for the U.S., relative to all 191 countries in the sample.

Figure 3 shows a comparison of all order-α efficiency rankings in 1997, including those found in Greene (2004) and Evans et al. (2000), without the inclusion of the percent of public expenditures that are publicly financed as an input variable.

Figure 3. Comparison of hyperbolic order-α efficiency rankings in 1997 across models



Note: Figure 3 shows the scatterplots of efficiency rankings from one model against another model. The upper-diagonal of the graph shows the correlation coefficients from the scatterplots; each of these compares ranking correlations for a pair of models. M*i* refers to the *i*th model described in Table 1, while Greene and Evans refer to the rankings found in the original analyses of Greene (2004) and Evans et al. (2000). The function of best fit is plotted with the scatterpoints.

All efficiency rankings are relative to only OECD countries. The upper-diagonal of Figure 3 shows the correlation coefficient between efficiency rankings in models. Figure 3 shows that there is a general upward trend in efficiency rankings. More efficient countries from one model tend to be more efficient countries in another model. The spread of individual scatter plots is high, which does suggest that there is still quite a high degree of variation in rankings across models, especially the scatter plots comparing my results with those of Greene (2004) and Evans et al. (2000). This highlights a common finding in the literature;

that efficiency rankings across heterogeneous countries can lead to less robust efficiency rankings. Comparisons of efficiency rankings between models that only use OECD countries as their sample set, Models 1, 2, and 5, have less variation in the scatter plots, highlighting the robustness across models.

In Table 8 I present the Malmquist index for the United States using DALY and ISR as outputs, between the years 1997 and 2005.

Table 8. Hyperbolic Malmquist index, OECD dataset, 1997/2005.

	Individual component of the Malmquist index				
Country	M	PE	S	PT	ST
Australia	-	-	-	-	-
Austria	1.0391	0.9999	0.9420	1.0078	1.0947
Belgium	-	-	-	-	-
Canada	-	-	-	-	-
Czech Republic	1.1087	0.9996	1.1136	0.9999	0.9961
Denmark	1.0058	1.0002	0.9116	0.9997	1.1035
Finland	1.0358	1.0000	0.9389	1.0024	1.1005
France	1.0699	1.0007	0.9691	1.0070	1.0955
Germany	1.1320	1.0008	1.0253	1.0004	1.1027
Greece	1.0708	0.9999	0.9795	1.0081	1.0846
Hungary	1.1962	0.9981	1.1550	1.0403	0.9975
Iceland	1.0469	1.0000	0.9490	0.9993	1.1039
Ireland	1.0527	0.9999	0.9851	0.9997	1.0691
Italy	-	-	-	-	-
Japan	1.0356	1.0000	0.9430	1.0008	1.0972
Luxembourg	1.0285	0.9999	0.9324	1.0008	1.1023
Mexico	-	-	-	-	-
Netherlands	1.0132	1.0013	0.9173	0.9988	1.1045
New Zealand	1.0611	0.9998	1.0070	0.9995	1.0545
Norway	1.0551	1.0003	0.9562	0.9988	1.1045
Poland	-	-	-	-	-
Portugal	-	-	-	-	-
Slovakia	1.2128	1.0022	1.1392	1.1385	0.9331
South Korea	1.2060	1.0000	1.1491	1.0811	0.9708
Spain	1.0904	1.0000	0.9949	1.0136	1.0812
Sweden	1.0309	1.0000	0.9345	0.9989	1.1043
Switzerland	1.0004	0.9993	0.9075	0.9974	1.1060
Turkey	-	-	-	-	-
UK	1.0413	0.9991	0.9448	1.0109	1.0912
US	0.9919	1.0010	0.8983	0.9987	1.1046

Notes: The Malmquist index (M) is defined for the OECD dataset (Model 5), between the years 1997 and 2005, using as output measures infant survival rate (ISR) and disability adjusted life years (DALY). PE refers to pure efficiency; S refers to scale; PT refers to pure technology; ST refers to scale technology. A value of less than 1 means that productivity has improved, while a number greater than 1 means that productivity has regressed.

A value of less than 1 indicates an improvement in productivity, while a value of greater than 1 indicates a decline in productivity. The results from the Malmquist scores I obtain support the findings from Hollingsworth and Wildman (2003), who observed that productivity regressed during the years 1993 to 1997. For all OECD countries observed but the United States, productivity regressed from 1997 until 2005. Interestingly enough, even though the U.S. is one of the more inefficient healthcare producing countries, it showed no statistically significant change in productivity over this time period. Overall, consistent with Hollingsworth and Wildman's (2003) findings, productivity regressed across the entire panel of OECD countries by around 5%. There seem to be several potential explanations for this phenomenon.

The first is the age demographics in the countries. Due to the baby boomer generation, there is a much larger percentage of the population that is older. The marginal life expectancy for an additional dollar in expenditures is much lower than what you would see in a younger population. The second is the increased spending on end-of-life care in developed countries. In many cases, end-of-life medical procedures and care that are proposed are costly and will not significantly improve a country's health outcomes. It seems as if economic considerations of cost-benefit analyses are not met, meaning more weight is attached to the moral imperatives of end-of-life care. A third explanation is that there have been no large medical technology breakthroughs during the observed time period (like penicillin or open heart surgery) that significantly increase health outcomes, while providing cheaper alternatives (and sometimes replacing more expensive existing technologies).[19] A fourth explanation is the shift from preventive medicine (exercise, diets, etc.) to reactive medicine (medical care after acquiring an illness/injury, etc.) due to improved reactive medical technologies. This leads to worse short- and long-term health outcomes, which means higher future health care costs. All of these seem to be plausible explanations for the productivity regression that this study has shown. A problem is that the higher health care costs do not seem to be compensated for by markedly better health outcomes.

An interesting result seems to be the amount of productivity regression in

[19] Advances in cancer and AIDS research, along with the introduction of genomic testing, may be a sufficient technological advance in the coming years to forestall some of this productivity regression.

Eastern European countries that used to be part of the Soviet bloc (Hungary, Czech Republic, and Slovakia). Hungary experienced regression of almost 20-percent, while Slovakia experienced regression of 21-percent. Part of this inefficiency may have occurred due to the switch from a Communist-based economy to a less restrictive economic regime. This may highlight the high levels of productive inefficiency that are caused by significant structural changes within a country. This result may be useful when examining any proposed changes to the health care delivery system in the United States.

V. Conclusion

I examine the technical efficiency of health care delivery systems in OECD countries based on two datasets. Previous studies, using older non-parametric estimators, suffer from method-related limitations. The DEA and FDH estimators suffer from the curse of dimensionality and other data-related problems. These limitations are partially alleviated using newer non-parametric estimators, like the order-α estimator and the order-m estimator. Likewise, while they estimate both input-oriented and output-oriented efficiency measures, newly developed hyperbolic-oriented efficiency measures take into account both of these margins, highlighting that they may be superior. The estimates do support what seems to be a common finding; that the U.S. healthcare delivery system is one of the more inefficient systems found in the world. Similarly, my estimates provide evidence that the newer efficiency estimates are robust across specifications with mild changes in outputs used when looking at OECD countries only, holding the estimator and input mix fixed. This leads to an important conclusion; as non-parametric estimation methods suffer from the limitations of not being able to remove cross-unit heterogeneity as well as parametric estimators, limiting the sample set so that it is more homogeneous can alleviate many of these problems. One such grouping, as I have found in my paper, is comparing OECD countries with each other.

These results suggest that removing as much cross-unit heterogeneity from non-parametric efficiency estimates as possible will continue to increase the robustness of the results, as this heterogeneity can only fully be removed in a second-stage formulation.[20] This suggests that within country efficiency estimates may be a superior option in determining the impacts of healthcare inputs on healthcare outputs. Though there may be heterogeneity within a country, it will be less than the heterogeneity of political systems, cultures, and behaviors in different countries. This suggests that efficiency rankings may be more robust when looking within a country, and may provide a more stable platform on which to build policy recommendations to alter the health care delivery system.[21] It is likely that adopting specific programs implemented in other states will yield results with minimal unintended consequences, when compared to adopting programs from other countries. This suggests further research: to address the notion that health care in the United States is uniquely inefficient, by looking at American health care efficiency at the state level.

I also find that the inclusion of questionable input variables, such as the percent of healthcare expenditures that are publicly financed, can drastically alter the efficiency rankings. The U.S. moves from dead last, in terms of efficiency rankings among the OECD countries, to the top quarter of healthcare delivery systems in the world (and, in one specification, is the most efficient producer of healthcare). This suggests another limitation of cross-country efficiency comparisons in healthcare; the inclusion of additional variables can introduce large amounts of bias in the estimates. Policymakers can specification search, by adding (or subtracting) input or output variables to obtain the desired results. The addition of simple, widely collected variables can have drastic effects on the efficiency estimates. Research should address whether these input variables are more appropriately used as an environmental variable in a two-stage regression, despite the known limitations of this type of analysis, and what impacts these environmental variables can have.

I find that there was general productivity regression across all of the countries observed, from 1997 to 2005, except for the United States, which is found to have no productivity change over the relevant time period. I also find that former Soviet countries (Hungary, Czech Republic, Slovakia) showed a significant

[20] Squires (2011) suggests that cross-national health care comparisons can guide policy. He uses a variety of summary statistics to compare health outcomes and resources used to deliver health care within a country.

[21] In other words, if efficiency rankings within a country are more robust to changes in input-output bundles and estimators, this provides more validity for the rankings and indicates where policy makers should begin to address changes in the underlying health care delivery system.

decline in productivity over this time period. Further research should explore the causes behind this general worldwide decline in productivity, and the apparent contradictory result for the U.S. It may also hint at general trends in the U.S. with the reform introduced with the Affordable Care Act (ACA); declines in productivity may not be attributable to the ACA, but instead attributable to general worldwide trends mentioned in the Results section, such as an aging (and unhealthy) population.

Further research seems to point to the need for a study of efficiency within a country to reduce the variation found even when comparing the healthcare systems of OECD countries with one another. Such research can then be used to provide policy recommendations, to pinpoint geographical areas or types of institutions (i.e., types of hospitals) that seem to be inefficient. Conducting efficiency measurements within a country also holds more things constant than cross-country comparisons, as even different areas within a country are likely to have populations that are, on average, more homogeneous than populations between countries. Similarly, research should focus on robustness checks of the wide variation in healthcare input-output combinations used in the efficiency literature across countries. As seen before, using the percent of healthcare expenditures that are publicly financed as an input (rather than per capita healthcare expenditures) leads to drastically different efficiency rankings. The use of a wide variety of financial and social input and output indicators calls into question the reliability of much of this literature.

References

Adam, Antonis, Manthos Delis, and Pantelis Kammas (2011). Public sector efficiency: leveling the playing field between OECD countries. *Public Choice* 146: 163-183.

Afonso, António, and Miguel St. Aubyn (2005). Non-parametric approaches to education and health efficiency in OECD countries. *Journal of Applied Economics* 8: 227-246.

Aigner, Dennis, C.A. Knox Lovell, and Peter Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6: 21-37.

Allen, Rachel, Antreas D. Athanassoupoulos, Robert G. Dyson, and Emmanuel Thanassoulis (1997). Weight restrictions and value judgments in data envelopment analysis: evolution, development, and future direction. *Annals of Operations Research* 73: 13-34.

Aragon, Yves, Abdelaati Daouia, and Christine M. Thomas-Agnan (2005). Nonparametric frontier estimation: a conditional quantile-based approach. *Econometric Theory* 21: 358-389.

Berger, Mark, and Jodi Messer (2002). Public financing of health expenditures, insurance, and health outcomes. *Applied Economics* 34: 2105-2113.

Cazals, Catherine, Jean-Pierre Florens, and Léopold Simar (2002). Nonparametric frontier estimation: a robust approach. *Journal of Econometrics* 106: 1-25.

Charnes, Abraham, William W. Cooper, and E. Rhodes (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429-444.

Coelli, Timothy J., D.S. Prasada Rao, Chris J. O'Donnell, and George E. Battese (2005). Data envelopment analysis. In *An introduction to efficiency and productivity analysis*, 2nd edition. New York, NY, Springer Science.

Daouia, Abdelaati, and Léopold Simar (2007). Nonparametric efficiency analysis: a multivariate conditional quantile approach. *Journal of Econometrics* 140: 375-400.

Dyson, Robert G., Rachel Allen, Ana S. Camanho, Victor V. Podinovski, Cláudia S. Sarrico, and Estelle A. Shale (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research* 132: 245-259.

Evans, David B., Ajay Tandon, Christopher J.L. Murray, and Jeremey A. Lauer (2000). The comparative efficiency of national health systems in producing health: an analysis of 191 countries. *GPE Discussion Paper 29*, World Health Organisation.

Garber, Alan M., and Jonathan Skinner (2008). Is American health care uniquely inefficient? *Journal of Economic Perspectives* 22: 27-50.

Greene, William (2004). Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems. *Health Economics* 13: 959-980.

Hollingsworth, Bruce, and John Wildman (2003). The efficiency of health production: re-estimating the WHO panel data using parametric and non-parametric approaches to provide additional information. *Health Economics* 12: 493-504.

Kim, Younhee, and Minah Kang (2014). The measurement of healthcare efficiency: cross-country comparison by geographical region. *The Korean Journal of Policy Studies* 29: 21-44.

Kneip, Alois, Byeong Park, and Léopold Simar (1998). A note on the convergence of non-parametric DEA efficiency measures. *Econometric Theory* 14: 783-793.

MacDorman, Marian F., and T.J. Mathews (2009). Behind international rankings of infant mortality: how the United States compares with Europe. *NCHS Data Brief 23*, National Center for Health Statistics.

Mathers, Colin, Ritu Sadana, Joshua Salomon, Christopher J.L. Murray, and Alan Lopez (2000). Estimates of DALE for 191 countries: methods and results. *GPE Discussion Paper 16*, World Health Organization.

Meeusen, Wim, and Julien van den Broeck (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review* 18: 435-444.

National Research Council and Institute of Medicine (2013). U.S. Health in international perspective: shorter lives, poorer health. Washington, DC, The National Academies Press.

Pipes, Sally (2013). Those misleading world health rankings. *The Wall Street Journal*.

Richardson, Jeff, John Wildman, and Iain Robertson (2003). A critique of the World Health Organization's evaluation of health systems performance. *Health Economics* 12: 355-366.

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

Simar, Léopold, and Paul Wilson (1999). Estimating and bootstrapping Mamlquist Indices. *European Journal of Operational Research* 115: 459-471.

Simar, Léopold, and Paul Wilson (2008). Statistical inference in non-parametric frontier models: recent developments and perspectives. In H. O. Fried, C. A. K. Lovell, S. S. Schmidt, editors, *The measurement of productive efficiency and productivity growth*. Oxford, Oxford University Press.

Squires, David (2011). The U.S. health system in perspective: a comparison of twelve industrialized nations. The Commonwealth Fund.

Tandon, Ajay, Christopher J.L. Murray, Jeremy A. Lauer, and David B. Evans (2000). Measuring overall health system performance for 191 countries. *GPE Discussion Paper 30*, Geneva, World Health Organisation.

Wheelock, David, and Paul Wilson (1999). Technical progress, inefficiency, and productivity changes in US banking, 1984-93. *Journal of Money, Credit, and Banking* 31: 212-234.

Wheelock, David, and Paul Wilson (2008). Non-parametric, unconditional quantile estimation for efficiency analysis with an application to Federal Reserve check processing operations. *Journal of Econometrics* 145: 209-225.

Wheelock, David, and Paul Wilson (2009). Robust non-parametric quantile estimation of efficiency and productivity changes in U.S. commercial banking, 1985-2004. *Journal of Business and Economic Statistics* 27: 354-368.

Williams, Alan (2001). Science or marketing at WHO? A commentary on 'World Health 2000'. *Health Economics* 10: 93-100.

Wilson, Paul (2011). Asymptotic properties of some non-parametric hyperbolic efficiency estimators. In I. van Keilegom and P. W. Wilson*, editors, Exploring research frontiers in contemporary statistics and econometrics*. Berlin, Springer-Verlag.

World Health Organisation (2000). *The World Health Report 2000 - Health systems: improving performance*. *Geneva,* World Health Organisation.