PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Measuring the efficiency of health systems in Asia: A data envelopment analysis
AUTHORS	Ahmed, Sayem; Hasan, Md. Zahid; MacLennan, Mary; Dorin, Farzana; Ahmed, Mohammad; Hasan, Md. Mehedi; Hasan, Shaikh Mehdi; ISLAM, Mohammad Touhidul; Khan, Jahangir

VERSION 1 – REVIEW

REVIEWER	Hong Wang BMGF, USA
REVIEW RETURNED	27-Apr-2018

GENERAL COMMENTS	This paper was trying to address a very important policy issue. However, the results were not so significant from policy perspective. Unlike the paper cited that there were about 20-40% of inefficiency in the system performance, this analysis was only able to capture 2%-7% of inefficiency of system performance, which is only a very small fraction of the overall inefficiency loss. Policymakers might not be motivated to take great efforts if they are only able to increase 2% technical efficiency in its healthcare system. In addition, the “output” variables selected in the paper is really the “outcome” variables. I was wondering if the analysis can use true health system “output” variables, the results will make the differences or not. A minor potential error in line 216, the number of the countries should be 46 instead of 460, correct?

REVIEWER	Michel Grignon McMaster University, Canada
REVIEW RETURNED	04-Jul-2018

GENERAL COMMENTS	This is a DEA of technical (and, in a way, cost) efficiency of health care systems in 46 countries of the Asian continent. Outputs are (inverse) infant mortality and life expectancy at birth and inputs are health care spending per capita in the country (in 2011 $PPP), beds and physicians densities (two variables), smoking prevalence (inverse?) and primary school completion rates in the relevant age group (?). One third of Asian countries are on the frontier and average inefficiency (waste) is relatively small (this comment is in regards to the VRS model). Low income Asian countries could improve their outcomes by 7% without spending more on health care or adding resources to their health care systems. Data are from the World Development Indicators database, wave 2014

(except when data were missing for a particular country and a particular year, in which case the most recent data before 2014 was used instead).

Overall, I find this manuscript unconvincing:
1) The background section does not really convince me that: a) this is the first ever study to measure technical efficiency of health care systems for all Asian countries and, b) that it is a study worth conducting. Such a measurement is of interest if it has real policy implications, meaning that the DMUs truly belong to one common production function and that they can learn from each other. I am not sure Afghanistan can learn from Japan, and a good background section should convince me that this is actually the case. Another possible justification for such a study would be that Asian health systems are very different from those in the rest of the world and we could learn something on efficiency in health care that we did not know without studying Asian health systems. I doubt it, but the case could be made for it if it were true.
2) The selection of inputs and outputs is not really discussed, except to write that these are variables commonly used in similar studies, or, later in the manuscript, that it was based on data availability. I don't agree with such a justification: the fact that a variable is available in a data set does not make it a good candidate for a study, and the fact that a variable is not in a data set does not rule it out as a good candidate (effort should be put in creating a proxy based on other variables).
1. Outputs: How do we know that health care systems are supposed to maximize life expectancy at birth and minimize infant mortality? How do we know that all systems/countries/leaders pursue the same objective? Some would argue that a health care system's mission is to make sure that individuals get timely access to care that is needed; this would point toward Potential Years of Life Lost as a better indicator of output. It could also be argued that quality of life matters as much as, if not more than, quantity of life and life expectancy should be weighted by disability or any measure of health. It would be unfair (and useless) to assess a country with one of these objectives on the basis of an objective they do not pursue. Last, it looks like infant mortality is accounted twice in this study: life expectancy at birth includes (and depends heavily on) survival from birth to 1st birthday. I would prefer life expectancy at age 1 (or any age greater than 1) and inverse infant mortality as the two outputs.
2. Inputs: My issue here is with having three very different types of "inputs" in the analysis: level of spending is well suited to a cost-efficiency analysis (rather than technical), as it mixes together the volume of resources and their costs. Beds and physicians densities are of the pure technical efficiency type (how many empty beds or idle doctors?) but could be complemented with nurses densities, as having too many nurses per doctor might be a source of inefficiency as well. Last, smoking and education are environmental variables: health care systems cannot really increase or decrease education rates, and they can only marginally influence smoking. I fully agree that these two variables affect efficiency (it is certainly easier to get good results per unit of input among a highly educated population), but am reluctant treating them as inputs, especially when tests of scale efficiency are implemented: if a health system is "too large", should it also increase smoking prevalence or reduce primary school completion? Here I make the assumption that the actual input for smoking is 100-prevalence rate, rather than prevalence rate, but it

	is not told in the manuscript. My preference regarding inputs would be to use spending only (cost-efficiency analysis) and then use beds and physicians (and nurses) densities as well as smoking prevalence and education as factors in a regression explaining the efficiency scores estimated in the first step. 3) Method: as briefly alluded to in the discussion section, DEA is a deterministic method (contrary to SFA) and it is highly vulnerable to outliers (mostly high achievers). One way around this is to use the bootstrap method developed by Simar and Wilson (1998) ("Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models", Management Science 44(1)). This is an absolute requirement in two-step analyses (calculate scores, then explain scores by their determinants), to make sure the series is not serially correlated (one score does not influence other scores) and it is greatly advised in single-step analyses. It usually reduces the number of DMUs on the frontier and lowers average efficiency scores. 4) Beside bootstrapping, a thorough sensitivity analysis is required in all DEA analyses: standard robustness checks include using a different year (not too far apart), excluding some high performers (outliers in life expectancy, like Japan, or in spending, like Bangladesh), and trying various measures of key variables (e.g., various PPP standardizations of the spending variable), as well as various measures of output (e.g., LE at age 60, PYLL etc.) 5) Segmenting findings by levels of income is certainly of great interest (to be honest, scores per se are not that interesting, but knowing how scores vary across characteristics of the country provides real insights into the question of the determinants of efficiency), but one wonders why income only. Why not segment by population density or share or rural inhabitants in the population, or political variables (free elections or not)? Which brings us back to the idea of a two-step analysis. 6) The manuscript should be clearer on some key aspects: for instance, what is the "relevant" age group on which education is measured?

REVIEWER	Li Wang Offord Center of Child Health Study, McMaster University, Canada
REVIEW RETURNED	05-Jul-2018

GENERAL COMMENTS	Comments: First off, I would like to thank the authors for all the hard work that they have put into this article submission. I found the paper to be interesting, particularly the comparison at the macro level. Constructively, I would like to highlight a few points which I believe could help improve the submission. Background: 1. I feel the motivation of the study is rather poor and must be improved. Is anything related to the policy motivation? 2. In the line 15, the author mentioned a few studies on the health systems efficiency across Asian countries. What are their main findings? The contribution to the existing literature need to be well defined. 3. In the line 103, the paper also evaluates the scale efficiency of the healthcare system. Methods: Inputs and output variables 1. In the line 111, just want to confirm that the number of beds and the physician cost are not included in the total expenditure?

	2. In the line 114, I think the reason of including the health status is to take into account the need of health care. Need to clarify why two environment variables are included as inputs. Data: 1. The imputation approach: If the historical data is available, is there time trend of the selected variables from the historical data? If yes, the time trend needs to be considered for the imputation. 2. Sensitivity analysis needs to conduct to examine the difference in the efficiency when using the completed data (the small number of countries). Data Envelopment Analysis 1. In the line 165, please specify why use the output orientation in this study. 2. Ratio variables only work in the CRS. Please explain why the ration inputs are ok in VRS. Check the reference: Hollingsworth, B & Smith, P 2003, 'The use of ratios in data envelopment analysis' Applied Economics Letters, vol 10, no. 11, pp. 733-735. 3. The limitation of the DEA in this study? Did you try the bootstrapping to get the CI of the point estimation of the efficiency score? 4. In the most of studies, two-stage DEA is the common method, i.e., to evaluate the efficiency score at the first stage, and the Tobit Regression used at the second stage to identify the determinants of the efficiency. I wonder why the study only have the first stage. Results 1. It would be interesting to check the correlation between the efficiency scores and the input variables. 2. the scale inefficiency is not well interpreted. 3. Did you use any statistical test to examine the differences across the income categories of the countries? Discussion About the discussion, I believe that it lacks some policy and managerial implications. Why did you perform your research if no clear implications can be drawn?

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
Reviewer Name: Hong Wang
Institution and Country: BMGF, USA
Competing Interests: None

Comment 1. This paper was trying to address a very important policy issue. However, the results were not so significant from policy perspective. Unlike the paper cited that there were about 20-40% of inefficiency in the system performance, this analysis was only able to capture 2%-7% of inefficiency of system performance, which is only a very small fraction of the overall inefficiency loss. Policymakers might not be motivated to take great efforts if they are only able to increase 2% technical efficiency in its healthcare system.
Response:
We thank the reviewer for the appreciation. Asian countries are not homogenous in terms of area, population, and economic conditions, however, they have public health functions and a number of their health system outcomes in common. Many of the countries share similar health systems

problems, including inadequate resources for healthcare and a high burden of diseases due to the geographical contiguity, disease patterns, and social conditions. Understanding health systems efficiency in different Asian countries could promote shared learning and highlight key areas of best practice, as well as areas where improvement is needed. Furthermore, given geographical proximity and many strong relationships experienced with near-by countries, there is likely to be relative ease in the ability to practically understand, learn and apply nuance about healthcare systems from one country to another. A study of the efficiency of health systems in this region will help to provide lessons through comparison across countries.

Comment 2. In addition, the "output" variables selected in the paper is really the "outcome" variables. I was wondering if the analysis can use true health system "output" variables, the results will make the differences or not.
Response: We agree with the reviewer, in the DEA model the variables are treated as input and output variables. The goal of a health system may be different for different countries, however, the status of a health system is measured by outcome e.g. mortality, life expectancy), in this study we used health systems outcome as the production function variables and we have revised in the manuscript output to the outcome.

Comment 3. A minor potential error in line 216, the number of the countries should be 46 instead of 460, correct?
Response: Thank you for this correction, the total country studied is 46.

Reviewer: 2
Reviewer Name: Michel Grignon
Institution and Country: McMaster University, Canada
Competing Interests: None declared.

This is a DEA of technical (and, in a way, cost) efficiency of health care systems in 46 countries of the Asian continent. Outputs are (inverse) infant mortality and life expectancy at birth and inputs are health care spending per capita in the country (in 2011 $PPP), beds and physicians densities (two variables), smoking prevalence (inverse?) and primary school completion rates in the relevant age group (?). One third of Asian countries are on the frontier and average inefficiency (waste) is relatively small (this comment is in regards to the VRS model). Low income Asian countries could improve their outcomes by 7% without spending more on health care or adding resources to their health care systems. Data are from the World Development Indicators database, wave 2014 (except when data were missing for a particular country and a particular year, in which case the most recent data before 2014 was used instead).

Overall, I find this manuscript unconvincing:
Comment 1. The background section does not really convince me that: a) this is the first ever study to measure technical efficiency of health care systems for all Asian countries and, b) that it is a study worth conducting. Such a measurement is of interest if it has real policy implications, meaning that the DMUs truly belong to one common production function and that they can learn from each other. I am not sure Afghanistan can learn from Japan, and a good background section should convince me that this is actually the case. Another possible justification for such a study would be that Asian health systems are very different from those in the rest of the world and we could learn something on efficiency in health care that we did not know without studying Asian health systems. I doubt it, but the case could be made for it if it were true.

Response: We have revised the background and updated the motivation for this paper and added the motivation as follows “Asian countries are not homogenous in terms of area, population, and economic conditions, however, they have public health functions and a number of their health system outcomes in common. Many of the countries share similar health systems problems, including inadequate resources for healthcare and a high burden of diseases due to the geographical contiguity, disease patterns, and social conditions. Understanding health systems efficiency in different Asian countries could promote shared learning and highlight key areas of best practice, as well as areas where improvement is needed. Furthermore, given geographical proximity and many strong relationships experienced with near-by countries, there is likely to be relative ease in the ability to practically understand, learn and apply nuance about healthcare systems from one country to another. A study of the efficiency of health systems in this region will help to provide lessons through comparison across countries.” (page 4, para-first)

Comment 2. The selection of inputs and outputs is not really discussed, except to write that these are variables commonly used in similar studies, or, later in the manuscript, that it was based on data availability. I don’t agree with such a justification: the fact that a variable is available in a data set does not make it a good candidate for a study, and the fact that a variable is not in a data set does not rule it out as a good candidate (effort should be put in creating a proxy based on other variables).
Response: We have now discussed the selection of inputs and outcome variables in this version. We selected the input variables as proxies for the quantity of inputs that a country devotes to healthcare (i.e. health expenditure per capita); and outcome variables of healthy life expectancy (HALE) at birth and infant mortality (per 1,000 live births). The relationship between health expenditure and outcomes considered here is consistent with the view that health expenditure has diminishing returns, or additional expenditure beyond a certain level has relatively smaller incremental effect on life expectancy or infant mortality (Morris et al. 2012). To be clear, reduction in infant mortality and increase in life expectancy signify improvement in the health outcomes of a country. Some studies have included life expectancy at birth as an outcome variable (Kirigia et al. 2011; Retzlaff-Roberts, Chang, and Rubin 2004; Wranik 2012), however, it is argued that quality of life matters as much as, if not more than, quantity of life, and therefore life expectancy should be a weighted health quality measure. As a result, HALE has been incorporated as a proxy of health quality as the outcome of health systems. Also, it is important to note that instead of using the infant mortality directly in the DEA model, we used the inverse of infant mortality as the model assumes that inputs and outputs are isotonic (i.e. increased input reduces efficiency as well as increased output increases efficiency) (Spinks and Hollingsworth 2009). Without this correction, a higher infant mortality figure would have been said to incorrectly contribute to a better health system outcome. (page 4-last para)

Comment 3. Outputs: How do we know that health care systems are supposed to maximize life expectancy at birth and minimize infant mortality? How do we know that all systems/countries/leaders pursue the same objective? Some would argue that a health care system’s mission is to make sure that individuals get timely access to care that is needed; this would point toward Potential Years of Life Lost as a better indicator of output. It could also be argued that quality of life matters as much as, if not more than, quantity of life and life expectancy should be weighted by disability or any measure of health. It would be unfair (and useless) to assess a country with one of these objectives on the basis of an objective they do not pursue. Last, it looks like infant mortality is accounted twice in this study: life expectancy at birth includes (and depends heavily on) survival from birth to 1st birthday. I would prefer life expectancy at age 1 (or any age greater than 1) and inverse infant mortality as the two outputs.
Response: We agree with the reviewer that the health system does not only focus on the quantity of life rather quality matters. We have revised out analysis considering the input as health expenditure per capita and outcome variable as healthy life expectancy (HALE) at birth and inverse infant mortality. The revised result is presented in table 2

Comment 4. Inputs: My issue here is with having three very different types of “inputs” in the analysis: level of spending is well suited to a cost-efficiency analysis (rather than technical), as it mixes together the volume of resources and their costs. Beds and physicians densities are of the pure technical efficiency type (how many empty beds or idle doctors?) but could be complemented with nurses densities, as having too many nurses per doctor might be a source of inefficiency as well. Last, smoking and education are environmental variables: health care systems cannot really increase or decrease education rates, and they can only marginally influence smoking. I fully agree that these two variables affect efficiency (it is certainly easier to get good results per unit of input among a highly educated population), but am reluctant treating them as inputs, especially when tests of scale efficiency are implemented: if a health system is “too large”, should it also increase smoking prevalence or reduce primary school completion? Here I make the assumption that the actual input for smoking is 100-prevalence rate, rather than prevalence rate, but it is not told in the manuscript. My preference regarding inputs would be to use spending only (cost-efficiency analysis) and then use beds and physicians (and nurses) densities as well as smoking prevalence and education as factors in a regression explaining the efficiency scores estimated in the first step.
Response:
As suggested by “My preference regarding inputs would be to use spending only (cost-efficiency analysis) and then use beds and physicians (and nurses) densities as well as smoking prevalence and education as factors in a regression explaining the efficiency scores estimated in the first step.” We have accommodated this in our manuscript. The new analysis included the per capita health expenditure at PPP as the input, and healthy life expectancy at birth and inverse infant mortality as the outcome variables. To identify the determinants of the efficiency score, we included population density, physician density, beds density, smoking percentage, and primary completion rate of the relevant age group as the independent variables in the tobit model (table 3).

Comment 5. Method: as briefly alluded to in the discussion section, DEA is a deterministic method (contrary to SFA) and it is highly vulnerable to outliers (mostly high achievers). One way around this is to use the bootstrap method developed by Simar and Wilson (1998) (“Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models”, Management Science 44(1)). This is an absolute requirement in two-step analyses (calculate scores, then explain scores by their determinants), to make sure the series is not serially correlated (one score does not influence other scores) and it is greatly advised in single-step analyses. It usually reduces the number of DMUs on the frontier and lowers average efficiency scores.

Response:
We have conducted the bootstrap method suggested by Simar and Wilson using the “simarwilson” command in STATA. The findings of the bootstrap method is attached in the supplementary files. However, the findings of the bootstrap method were almost similar with the findings from the tobit regression method (supplementary table-2).

Comment 6. Beside bootstrapping, a thorough sensitivity analysis is required in all DEA analyses: standard robustness checks include using a different year (not too far apart), excluding some high performers (outliers in life expectancy, like Japan, or in spending, like Bangladesh), and trying various measures of key variables (e.g., various PPP standardizations of the spending variable), as well as various measures of output (e.g., LE at age 60, PYLL etc.)
Response:
Thank you for this important comment. We have now conducted sensitivity analysis using various combination of input and output variables, for instance, changing input, per capita health expenditure from international PPP USD to per capita health expenditure at current USD, removing the outliers i.e. efficient DMUs form the analysis, changing the output variables e.g. from healthy life expectancy at birth to health life expectancy at age 60, using complete set of data (excluding the countries with

missing variables). In all these cases the average of the efficiency scores varied from 0.812 to 0.936. The most sensitive case was found while considering the healthy life expectancy at age 60 as the input variable. The efficiency score changed from 0.919 (main model) to 0.812 (considering input as healthy life expectancy at age 60). We did not have scope to include the PYLL as the data were not available for Asian countries.

Comment 7. Segmenting findings by levels of income is certainly of great interest (to be honest, scores per se are not that interesting, but knowing how scores vary across characteristics of the country provides real insights into the question of the determinants of efficiency), but one wonders why income only. Why not segment by population density or share or rural inhabitants in the population, or political variables (free elections or not)? Which brings us back to the idea of a two-step analysis.

Response: We have now added the two-step analysis, second step as the tobit model. We have also added population density in addition with the income level of the countries as the determinants of health systems efficiency (table-3).

Comment 7. The manuscript should be clearer on some key aspects: for instance, what is the "relevant" age group on which education is measured?

Response: The relevant age group for the primary completion rate is defined as the number of new entrants (enrollments minus repeaters) in the last grade of primary education, regardless of age, divided by the population at the entrance age for the last grade of primary education of a country. The definition is added in the main text (page-8, para-last).

Reviewer: 3
Reviewer Name: Li Wang
Institution and Country: Offord Center of Child Health Study, McMaster University, Canada Competing Interests: None declared

Comments:
First off, I would like to thank the authors for all the hard work that they have put into this article submission. I found the paper to be interesting, particularly the comparison at the macro level.

Constructively, I would like to highlight a few points which I believe could help improve the submission.
Background:
Comment 1. I feel the motivation of the study is rather poor and must be improved. Is anything related to the policy motivation?
Response: We have now revised the section as follows "Asian countries are not homogenous in terms of area, population, and economic conditions, however, they have public health functions and a number of their health system outcomes in common. Many of the countries share similar health systems problems, including inadequate resources for healthcare and a high burden of diseases due to the geographical contiguity, disease patterns, and social conditions. Understanding health systems efficiency in different Asian countries could promote shared learning and highlight key areas of best practice, as well as areas where improvement is needed. Furthermore, given geographical proximity and many strong relationships experienced with near-by countries, there is likely to be relative ease in the ability to practically understand, learn and apply nuance about healthcare systems from one country to another. A study of the efficiency of health systems in this region will help to provide lessons through comparison across countries." (Page- 4, para-first)

Comment 2. In the line 15, the author mentioned a few studies on the health systems efficiency across Asian countries. What are their main findings? The contribution to the existing literature need to be well defined.
Response: We have now revised the section as follows "Findings from the existing literature is now added in the background section. A number of studies have addressed healthcare efficiency in Americas (Chattopadhyay & Ray, 1996; Shroff, Gulledge, Haynes, & O'Neill, 1998), Western Europe (Giuffrida & Gravelle, 2001; Hollingsworth & Parkin, 2001) and Asia (Chang, 1998; Wan et al., 2002) to shed light on the efficiency of different national health systems. A systematic review on measuring efficiency related to several aspects of healthcare was performed by Hollingsworth et al. (Hollingsworth, Dawson, & Maniadakis, 1999). Dimas et al. evaluated the productivity of Greek public hospitals and found that productivity changes were dominated by technical change (Dimas, Goula, & Soulis, 2012). Zere et al. measured the technical efficiency and productivity of hospitals in South Africa, and examined the impact of hospital characteristics on efficiency and productivity (Zere, Mcintyre, & Addison, 2005).
In an international study of efficiency in 170 countries, it was observed that Asian countries were comparatively in the middle with respect to health system efficiency scores (Kim & Kang, 2014). This indicates that there is room for improvement to optimize health benefits from the available health sector resources. In this region, there are a number of studies at the country level to address health systems efficiency (Cheng et al., 2015; Jat & Sebastian, 2013), but cross country comparison of the health system efficiency is limited (Hussey et al., 2009). (page-3, para- last)

Comment 3. In the line 103, the the scale efficiency of the healthcare system.
Response: We have revised now (Page-7, para-last).

Methods:
Inputs and output variables
Comment 4. In the line 111, just want to confirm that the number of beds and the physician cost are not included in the total expenditure?
Response: In the revised manuscript, we have considered Current health expenditure per capita (current US$) as the input variables. The other variables physician and beds density per 1000 population has been used in the tobit regression model to find the association with the efficiency scores (page 4 last para, page-8 first para).

Comment 5. In the line 114, I think the reason of including the health status is to take into account the need of health care. Need to clarify why two environment variables are included as inputs.

Data:
Response: We have excluded the two environmental variable as input variables and now these two have been used in the tobit model to as the determinant of efficiency (page-6, para-first).

Comment 6. The imputation approach: If the historical data is available, is there time trend of the selected variables from the historical data? If yes, the time trend needs to be considered for the imputation.
Response: Thank you for your suggestion. Although the variables used in the new DEA model was non-missing except for Syria. We have used the time trend for the missing variables for Syria. The other variables did not have time trend from the historical data and we could not use this method.

Comment 7. Sensitivity analysis needs to conduct to examine the difference in the efficiency when using the completed data (the small number of countries).
Data Envelopment Analysis

Response: We have conducted sensitivity analysis considering several factors; e.g. dropping the efficient countries, using healthy life expectancy at age 60, current health expenditure per capita (current US$) as inputs, and the completed data (the small number of countries). The result of the sensitivity analysis is presented in supplementary documents and also a figure is added (Figure 2).

Comment 8. In the line 165, please specify why use the output orientation in this study.

Response: In an input orientation DEA model the primary objective is to minimize the inputs used, whereas in an output orientation model, the objective is to attain the highest possible outputs with a given amounts of inputs. In this study, an output-oriented DEA model was deemed more appropriate based on the premise that per capita expenditure is essentially fixed inputs to work with at any given time. Alternatively, the health system stewards would have more leverage in controlling outputs through innovative programming rather than by spending more resources. (Page-6, para-last)

Comment 9. Ratio variables only work in the CRS. Please explain why the ration inputs are ok in VRS. Check the reference:
Hollingsworth, B & Smith, P 2003, 'The use of ratios in data envelopment analysis' Applied Economics Letters, vol 10, no. 11, pp. 733-735.

Response: The outcome of the health system as production function are variable returns to scale and follows the diminishing marginal returns. As a result, we have used the VRS DEA model. Moreover, a number of studies have followed the similar methods while analyzing the efficiency of health systems.

Comment 10. The limitation of the DEA in this study? Did you try the bootstrapping to get the CI of the point estimation of the efficiency score?
Response: Now we have added the results from bootstrapping of the efficiency scores in the table 4. After bootstrapping, the result was found to be similar with the tobit regression model. We have added the text as “The bootstrapping of VRS technical efficiency showed the similar association with as found using the tobit regression model” (supplementary table 2)

Comment 11. In the most of studies, two-stage DEA is the common method, i.e., to evaluate the efficiency score at the first stage, and the tobit Regression used at the second stage to identify the determinants of the efficiency. I wonder why the study only have the first stage.

Response: Thank you for this comment, we have now added the Tobit regression model as the second stage to identify the determinants of the efficiency. We have used bed density, physician density, percentage of male smokers, primary education completion rate of relevant age group, income category, and population density as the determinants of efficiency. We found that VRS efficiency was significantly associated with the bed density, primary education completion rate, and population density (number of population living per sq. kilometer land area). No significant association was observed with physician density and smoking prevalence.
(page-11, para-first)

Results
Comment 12. It would be interesting to check the correlation between the efficiency scores and the input variables.
Response: We have added the correlation between the efficiency scores and the input-output variables in the supplementary tables 6. Both VRS and CRS technical efficiency score were positively

correlated with per capita health expenditure, healthy life expectancy and negatively correlated with infant mortality STable 6 in the supplementary materials.

Comment 13. The scale inefficiency is not well interpreted.
Response: The scale efficiency is now interpreted as “Scale efficiency scores provide information on the optimality of the DMUs size. When a production unit (DMU) operates at CRS, TE is equal to scale efficiency. However, when DMUs are not operating at optimum scale, technical efficiency measured with the CCR model may be altered by scale efficiency. The BCC model, as the production is VRS, can incorporate the impact of scale efficiency in measurement of TE and this is measured as the ratio of CRS technical efficiency scores and VRS technical efficiency scores. (Page-5, para- last)

Comment 14. Did you use any statistical test to examine the differences across the income categories of the countries?
Response: No we did not use any statistical test to examine the differences across the income categories of the countries.

Discussion
Comment 15. About the discussion, I believe that it lacks some policy and managerial implications. Why did you perform your research if no clear implications can be drawn?

Response: The discussion is now revised.

VERSION 2 – REVIEW

REVIEWER	Li Wang McMaster University, Canada
REVIEW RETURNED	13-Sep-2018

GENERAL COMMENTS	1. Introduction: a. The author may want to address why you want to measure the scale efficiency across Asian countries. b. It may be helpful to address some important determinants of the efficiency from the current literature and what’s the policy implication. 2. Methods a. Why did the author use DEA and what’s the limitation of the DEA (e.g., random noise in the efficiency component, sensitive to the outliers)? How do you deal with this in your analysis? b. Input – It might be better to specify what include in the health expenditure? Do the components of the health expenditure are same across Asian countries? Did the capital values include in the health expenditure? c. Data envelopment analysis: It might be better to specify the means of the scale efficiency scores as well. d. Tobit regression analysis: i. Efficiency scores are relative values. The efficiency score has a serial correlation problem as the dependent variable in the regression. Why do not use “smoothed bootstrap” method designed by Simar and Wilson (1998,2007)? ii. Line 205: some true inputs of the health systems. – it might be better to use other words for “true inputs”. If they are regard as the input, do they double count in the evaluation of the efficiency?

	iii. It could regress on the efficiency to avoid transformation. iv. Line 245: “We considered multiple models (e.g. dropping the efficient countries, using HALE at age 60, current health expenditure per capita (current US$) as inputs”. -- The sensitivity analysis should specify clearly (eg., why do you want to test? How to test? What are you expected? 3. Results: a. Line 275: It’s better to interpret the efficiency score more understandable, e.g., does the 0.92 means the HALE would increase by 8%, it’s equivalent from the 64.29 to 64.29*1.08? Same for the scale efficiency. b. How much the variance could be explained by Tobit regression?

VERSION 2 – AUTHOR RESPONSE

Response to the reviewers' comments

Comment 1

1. Introduction:

a. The author may want to address why you want to measure the scale efficiency across Asian countries.

Response: We measured scale efficiency to see whether the health systems of Asian countries are operating at its’ optimal size or not. The size of health systems is a major political decision in Asian countries. To some extent, it depends on how the policymaker or government prioritizing health among other competing public services (e.g. education, military, electricity) (Achoki et al., 2017; WHO, 2010) (Page 6, para 2)

Comment 2

b. It may be helpful to address some important determinants of the efficiency from the current literature and what’s the policy implication.

Response: We have now added some determinants of the efficiency form the recent literature as follows. Several studies reported of different types of determinants to the health system efficiency. A study conducted in China reported that GDP per capita, proportion of primary health worker, and population density were the key determinants of the Chinese health system (Zhang et al., 2017). Another study reported that re-admission, obesity and smoking, and average income of the population are the key determinants of health system efficiency (Allin, Grignon, & Wang, 2014). (Page 3, 3rd para)

Comment 3

2. Methods

a. Why did the author use DEA and what's the limitation of the DEA (e.g., random noise in the efficiency component, sensitive to the outliers)? How do you deal with this in your analysis?

Response: Now we have added the limitation of DEA and used the bootstrap method in addition with the to address the random noise in the efficiency components. Following paragraph is now added in page 8, 2nd para.

"One of the limitation of the DEA approach is the serial correlation of the efficiency scores generated through this approach. The correlation between inputs and outputs, and consequently with the estimated efficiency scores resulted in this serial correlation. Thus, the scores of one DMU is not independent from that of the other DMUs. To handle this limitation, scholars such as Ramalho et al. 2010 (Ramalho, Ramalho, & Henriques, 2010) and McDonald 2009 (McDonald, 2009) have argued that econometric models like probit, logit, and truncated regression (Tobit) can be used for second-stage analysis for identifying impact of environmental variables on efficiency. However, scholars such as Simar and Wilson 2007 argued that the conventional statistical inferences are inappropriate in the second-stage regression due to the biasness of the DEA score and recommend use of bootstrap methods (Simar & Wilson, 2007). Afonso and Aubyn 2011 (Afonso & St. Aubyn, 2011) showed in their empirical study that the censored normal Tobit regression and bootstrap algorithms yielded very similar results. However, we have adopted the Tobit model and smoothed bootstrap model in explaining the association with health system efficiency."

Comment 4

b. Input – It might be better to specify what include in the health expenditure? Do the components of the health expenditure are same across Asian countries? Did the capital values include in the health expenditure?

Response: The health expenditure per capita was extracted from the Global Health Expenditure database managed by the WHO. In this database WHO maintaining national health expenditure statistics of more than 190 WHO Member States in line with the new System of Health Accounts 2011 (SHA 2011) framework. The SHA 2011 framework was developed by OECD to rigorously track health expenditure date (e.g. by all financial sources, by all services) at national level and maintaining comparability across the countries. The capital expenditure (e.g. infrastructure) was included in the total health expenditure estimation (31,32). (page 5, first para).

Comment 5

c. Data envelopment analysis: It might be better to specify the means of the scale efficiency scores as well.

Response: The average scale efficiency score was 0.847 (95% CI 0.824-0.87). We presented this in Table 2.

d. Tobit regression analysis:

i. Efficiency scores are relative values. The efficiency score has a serial correlation problem as the dependent variable in the regression. Why do not use “smoothed bootstrap” method designed by Simar and Wilson (1998,2007)?

Response: We have included the results from “smoothed bootstrap” analysis in Table 3.

Comment 6

ii. Line 205: some true inputs of the health systems. – it might be better to use other words for “true inputs”. If they are regard as the input, do they double count in the evaluation of the efficiency?

Response: We have now revised this sentence as “health service productions”. And they do not double count in the efficiency (page 8, last para).

Comment 7

iii. It could regress on the efficiency to avoid transformation.

Response: Now we have used the efficiency in the smoothed bootstrap regression model (Simar & Wilson, 2007) and transformation of the efficiency to a censored Tobit regression model. The negative association of explanatory variables with the transformed inefficiency score depicts the positive relation with the efficiency. And the positive association of the explanatory variables the with efficiency score in the smoothed bootstrap regression analysis depicts the positive relation with the efficiency scores (page 8, first para).

Comment 8

iv. Line 245: “We considered multiple models (e.g. dropping the efficient countries, using HALE at age 60, current health expenditure per capita (current US$) as inputs”. -- The sensitivity analysis should specify clearly (e.g., why do you want to test? How to test? What are you expected?

Response: We have now revised this section as suggested by the reviewer (page 10, second para).

Comment 9

3. Results:

a. Line 275: It’s better to interpret the efficiency score more understandable, e.g., does the 0.92 means the HALE would increase by 8%, it’s equivalent from the 64.29 to 64.29*1.08? Same for the scale efficiency.

Response: We have now revised the section as follows If all the health systems operated at maximum efficiency at their given input level, the high-, upper middle-, low- and lower-middle income countries could improve their health system outcome e.g. HALE at birth and reduce infant mortality by 6.6%, 8.7%, and 8.7% respectively. (page 14, first para)

Comment 10

b. How much the variance could be explained by Tobit regression?

Response: 23% variance was explained by the independent variables in the Tobit regression.

VERSION 3 – REVIEW

REVIEWER	Li Wang McMaster University
REVIEW RETURNED	20-Dec-2018

GENERAL COMMENTS	The paper is well revised and I recommend to accept. No further comments.