PEER REVIEW HISTORY

This paper was submitted to a another journal from BMJ but declined for publication following peer review. The authors addressed the reviewers' comments and submitted the revised paper to BMJ Open. The paper was subsequently accepted for publication at BMJ Open.

(This paper received three reviews from its previous journal but only two reviewers agreed to published their review.)

ARTICLE DETAILS

TITLE (PROVISIONAL)	Inequalities in healthcare resources and outcomes threatening sustainable health development in Ethiopia: Panel data analysis
AUTHORS	Woldemichael, Abraha; Takian, Amirhossein; Akbari Sari, Ali; Olyaeemanesh, Alireza

VERSION 1 – REVIEW

REVIEWER	Masatoshi Matsumoto Hiroshima University, Japan
REVIEW RETURNED	27-Mar-2018

GENERAL COMMENTS	This is an important study that evaluated the degree of inequity in health resource and government expenditure in Ethiopia. They seem to calculate Gini index and other equity indicators based on eleven regions in the country. I am concerned about the reliability of these indicators because the number of regions is too small. With only eleven geographic regions, I guess Lorenz curve is not really a curve but like stairs. In this context, how do you measure equity correctly? Authors stated that they estimated missing data on GHE and physicians, but I could not find how much data were missed in their data set. I think author should explain not only how they analysed their data but also how reliable the data were. Authors stated that the inequity of various resources and expenditure is high in Ethiopia. But the values of equity indicators vary according to how they demarcate geographic regions. So the values themselves cannot tell whether the distributions are equitable or not. Rather, comparison of the equity indicator values between different time points is much more meaningful. The results of this study showed a remarkable improvement of equity level in various resources and expenditure throughout the study period. You should focus more on this finding rather than the values themselves.

REVIEWER	Lingzhong Xu Department of Social Mediicne and Health Services Management, School of Public Helath, Shandong University, P.R.China
REVIEW RETURNED	07-Sep-2018

GENERAL COMMENTS	1.Line53-57,page 3, line1-22,page4: This part of the background should be appropriately streamlined. 2."0.000" is not an appropriate p-value (i.e., Table 3) - that value merely reflects the limitations of the statistical program. Such p-values are generally truncated to p<0.001 (as is written in the manuscript text). 3. All analyses were performed using STATA version 14. Please add company, city, state abbreviation. 4. Line 45, page 9: Amhara (0.385), This value does not correspond to the value in the Table 1. It should be 0.387. 5. Table 1 and Table 2: The brackets for the 95% confidence interval should be consistent in the English state, such as () or []. 6.There are also places where attention to scholarly writing is needed, for example, Line 51-54: The inequality in the GPs, all physicians, HOs, nurses plus midwives, “skilled health professionals”, and medical laboratory personnel distributions each reduced by more than a third ($p < 0.01$). Both the full name and the abbreviated name make the sentence confusing and difficult to understand. Therefore, all names of the indicators or regions are either the full name or the abbreviation, which should be consistent throughout the text and all table, such as Gini index (GI), Health officers (HOs),Somali (SO). 7. Some values in the text are not consistent with the values in the table, such as line35, page 12 and line 19, page 16 etc.I don't think there is a need for rounding. 8. Only 8.1% of references refer to the last triennium (2015-2018). There are several more recent studies in the subject.

REVIEWER	Marek Brabec Institute of Computer Science Czech Republic
REVIEW RETURNED	16-Oct-2018

GENERAL COMMENTS	This is a paper on interesting theme of substantial practical interest. Nevertheless, its present form is not suitable for publication. It should be substantially improved first. Below, there is a (partial) list of important changes to make for a possible re-submission. English of the paper should be improved substantially before publication. There are sentences lacking basic English structure in the text (see e.g. “The analysis was based publicly available secondary data for the regions.”, but also many other places). In the current form, the paper is very hard to read and sometimes the text is really non-decipherable. Sometimes, the problem is even at the level of really sloppy typography – like “… are skewed to Towards …”. Can you elaborate a bit more for the missing data limitation (mentioned on page 3)? Is it true that the missing data really bias the inequality estimate? Can you say that they bias it always downwards (in such a case, the estimates can be considered as being conservative)? Did you experiment with how much they can bias the results - either by simulating new data or labeling artificially some (randomly selected) data that were actually present as missing (and studying how much missings you can/cannot afford before spoiling the results completely)?

	The paragraph starting with “What are the trends and extent …” on page 5: would not it be better to speak about spatial trends, temporal trends and interaction (spatio-temporal inconsistencies – departures from smooth spatial+temporal trend model)? Do you rally want to say: “The ratios for were weighted by a fixed number of people for all the regions” (page 6) – which would amount to standardizing by a constant? Or, do you want to say that you acknowledged different number of people in different regions but did not have information about how the regional population changed over the study period, so that you used regional population counts from a single year (which one?)? Theil and Gini indexes should be defined (as a mathematical formula) on page 7. Statements like “The GI is one of the most commonly used measures of distributional inequalities in healthcare …” should be either supported (by direct evidence or citation of literature sources) or skipped. What do you mean by “This analysis utilises the value of observations greater than zero.” on page 8? Does it mean that you excluded 0 values from modeling? If yes, how much it smooths (decreases the inequality of) the results? Bootstrap and jackknife techniques mentioned on page 8 did not cope with the (potentially serious) effect of missing values, right? How exactly you fitted the multiplicative model on page 8? Did you take log of both sides, or did you even estimate different components on the right hand side separately? The statistical modeling and estimation procedures have to be described much more carefully and in substantially greater detail. This should include mathematical form of the model, assumption about error, concrete covariates and how they entered the model, estimation techniques used for identifying parameters etc. There are too many unnecessarily detailed tables that are hard to read. Could you present some of the results in the much-more-easily-comprehensible graphical form? The temporal trends (as overlay of curves for different regions etc.) are particularly suitable for presentation … The revised version of the paper should be substantially shortened and made much more compact.

REVIEWER	Tomoki Fujii Singapore Management University, Singapore
REVIEW RETURNED	21-Oct-2018

GENERAL COMMENTS	The description of the data is so vague that it is nearly impossible to understand what the authors did exactly. It appears to me from the description in p.5 lines 21-34 (line numbers refer to the numbers in the PDF) the authors have one data point per region per year. However, in page 26 (of 28), Gini is reported for each region. Therefore, the authors must have gotten some distributional information for each region, but it is unclear how it was obtained. I am also not sure whether the figures in this page

	were created from HHRIs or EDHS (a kind of information that should be included in the figure). Further, what is reported does not make too much sense to me. What is meant by "population ranked by share of "nurses + midwives"? What is top 10% in this context? Top 10% of what? How is Gini computed? This is reference to which year? There are so many obvious issues that need to be addressed. The authors make a very naive cross-country comparison. The level of measured inequality depends on the unit of analysis. Without addressing this issue, cross-country comparision in the avaialbility of healthcare personnel would be meaningless. Finally, the writing of the paper also needs to be substantially improvoed. There are obvious problems in writing (e.g., p.6 of 28, line 25. "The ratios for were weighted by a fixed number of people for all the regions." The sentence is grammatically incorrect. Also, what is meant by "fixed nubmer of people"?).

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

1. This is an important study that evaluated the degree of inequity in health resource and government expenditure in Ethiopia.

R: Thank you very much for your encouraging comments.

2. They seem to calculate Gini index and other equity indicators based on eleven regions in the country. I am concerned about the reliability of these indicators because the number of regions is too small. With only eleven geographic regions, I guess Lorenz curve is not really a curve but like stairs. In this context, how do you measure equity correctly?

R: We agree with this valuable suggestion. Although we included all the Regional States in Ethiopia (N = 11) as mentioned in the manuscript, the issue of small sample is evident. The regions also differ extremely in population size and there is a direct relationship between the Gini index/Lorenz curve and population size. A region with small population would have underestimated Gini index/Lorenz curves. While we included small sample bias correction term in the revised version of the manuscript, the small sample issue has been mentioned as a limitation of our study. Nevertheless, analysing inequalities using a small sample size has its roots in the literature, for example:

- *Alam N, Hajizadeh M, Dumont A, Fournier P. Inequalities in maternal health care utilization in sub-Saharan African countries: a multiyear and multi-country analysis. PloS one. 2015 Apr 8;10(4): e0120922.*
- *Li D, Zhou Z, Si Y, Xu Y, Shen C, Wang Y, Wang X. Unequal distribution of health human resource in mainland China: what are the determinants from a comprehensive perspective? International journal for equity in health. 2018 Dec;17(1):29.*

3. Authors stated that they estimated missing data on GHE and physicians, but I could not find how much data were missed in their data set. I think author should explain not only how they analysed their data but also how reliable the data were.

R: The missing data were on government health expenditure (GHE) (for the years 2013 and 2014) and on general practitioners (GPs) and specialist doctors (SPDs) (for the year 2015), which was stated on page 6, second paragraph of the “Variables (indicators)” sub-topic. Regarding the reliability of the data, as explained in the revised manuscript, this dataset is published and approved by the Federal Ministry of Health of Ethiopia, which to the best of our knowledge is the only one of its kind which is available.

4. Authors stated that the inequity of various resources and expenditure is high in Ethiopia. But the values of equity indicators vary according to how they demarcate geographic regions. So, the values themselves cannot tell whether the distributions are equitable or not. Rather, comparison of the equity indicator values between different time points is much more meaningful. The results of this study showed a remarkable improvement of equity level in various resources and expenditure throughout the study period. You should focus more on this finding rather than the values themselves.

R: Thank you for your comment. Our primary objective was to determine the aggregate (national) level inequalities over the 16 years period of the study. We also wanted to decompose the inequalities into components using the Pyatt's Gini decomposition technique (please refer Table 1 & 2). We used as an advantage to present the regional inequalities which were simultaneously generated using the same stata command. As endorsed by the reviewer, we focused on determining the changes in inequalities as indicated in Table 3 and 5.

Reviewer 2:

1. Line53-57, page 3, line1-22, page4: This part of the background should be appropriately streamlined.

R: Thank you for the constructive suggestion. The paragraph has been revised to accommodate the proposed changes.

2. "0.000" is not an appropriate p-value (i.e., Table 3) - that value merely reflects the limitations of the statistical program. Such p-values are generally truncated to p<0.001 (as is written in the manuscript text).

R: The necessary changes have been made and highlighted in yellow.

3. All analyses were performed using STATA version 14. Please add company, city, state abbreviation.

R: The suggestion has been incorporated

4. Line 45, page 9: Amhara (0.385), This value does not correspond to the value in the Table 1. It should be 0.387.

R: Thank you for providing us with another chance to check the possible errors. We have now corrected table.

5. Table 1 and Table 2: The brackets for the 95% confidence interval should be consistent in the English state, such as () or [].

R: The necessary changes have been made now

6. There are also places where attention to scholarly writing is needed, for example, Line 51-54: The inequality in the GPs, all physicians, HOs, nurses plus midwives, "skilled health professionals", and medical laboratory personnel distributions each reduced by more than a third ($p < 0.01$). Both the full name and the abbreviated name make the sentence confusing and difficult to understand. Therefore, all names of the indicators or regions are either the full name or the abbreviation, which should be consistent throughout the text and all table, such as Gini index (GI), Health officers (HOs), Somali (SO).

R: We have now made all the necessary changes throughout the manuscript.

7. Some values in the text are not consistent with the values in the table, such as line35, page 12 and line 19, page 16 etc.I don't think there is a need for rounding.

R: The inconsistencies have been identified and addressed accordingly

8. Only 8.1% of references refer to the last triennium (2015-2018). There are several more recent studies in the subject.

R: Some of the old references have been replaced by the latest ones. Out of the existing 86 reference, we remove 10 old references and added 15 references (12 latest and 3 old but relevant) references.

Reviewer 3:

1. This is a paper on interesting theme of substantial practical interest. Nevertheless, its present form is not suitable for publication. It should be substantially improved first. Below, there is a (partial) list of important changes to make for a possible re-submission.

R: Thank you very much for your valuable comments.

2. English of the paper should be improved substantially before publication. There are sentences lacking basic English structure in the text (see e.g. “The analysis was based publicly available secondary data for the regions.”, but also many other places). In the current form, the paper is very hard to read and sometimes the text is really non-decipherable.
Sometimes, the problem is even at the level of really sloppy typography – like “… are skewed to Towards …”.

R: We appreciate the comments raised for improving English in this manuscript. We have now revised the whole draft and had it proofread by an English native speaker. All the changes have been highlighted in the revised manuscript.

3. Can you elaborate a bit more for the missing data limitation (mentioned on page 3)? Is it true that the missing data really bias the inequality estimate? Can you say that they bias it always downwards (in such a case, the estimates can be considered as being conservative)? Did you experiment with how much they can bias the results - either by simulating new data or labeling artificially some (randomly selected) data that were actually present as missing (and studying how much missings you can/cannot afford before spoiling the results completely)?

R: Thank you for this comment. As it was partially explained in response to reviewer 1, we have also added further information concerning the missing data from the annually published bulletin. The missing data on GHE of two years (2013 and 2014) and on GPs and SPDs of one year (2015) does not imply that there was not health budget and physicians in the stated years. We noted up-ward biased Gini values when carrying out the analysis with the missing data for the concerned indicators at national and regional level. As it is well documented, using small sample, e.g., small units of analysis such as the regions in our case and sub-groups with different population size, will inevitably be underestimate the Gini values of the small sample or the unit with small population size in the analysis. It would also have created a problem in measuring the inequality change between two points in time. Thus, we have estimated the missing observations because they were not ignorable data and to handle the possible bias associated with the missing observations.

4. The paragraph starting with “What are the trends and extent …” on page 5: would not it be better to speak about spatial trends, temporal trends and interaction (spatio-temporal inconsistencies – departures from smooth spatial+temporal trend model)?

R: The regions in Ethiopia are structured based on ethnicity, except for AA and DD regions. As it was described in Fig.1 and Table 1 & 2, the agrarian regions in Ethiopia are dominantly the highlands and accounted for around 40% of the total land mass of Ethiopia. About 90% of the Ethiopians reside in the highland areas. In contrast, the low-land accounts for about 60% of the total land-mass, and nearly 10% of the Ethiopians (dominantly the pastoralists) inhabitant in this 60% of the total land mass. Therefore, we found it appropriate to use the population-based inequality analysis rather than the space-based inequality analysis.

5. Do you rally want to say: “The ratios for were weighted by a fixed number of people for all the regions” (page 6) – which would amount to standardizing by a constant? Or, do you want to say that

you acknowledged different number of people in different regions but did not have information about how the regional population changed over the study period, so that you used regional population counts from a single year (which one?)?

R: We had the CSA of Ethiopia census-based annual population estimates for each region for sixteen years (2000 to 2015). As we have presented it in Fig.1 of the manuscript, the regions widely differ in population size from 0.26% out of the total population of Ethiopia for HA region to 37.45% for OR region. The distributions we analysed also differ among the regions. However, an "indicator" should be consistent in measuring the ***equality or inequality*** of the distributions and the changes at a given time. If an indicator is not unique (uniform for a distribution across all the regions at a given time period), we cannot talk about inequality/disparity. Thus, the purpose of multiplying the ratio of a given distribution by a respective fixed number for all the regions was to ensure the consistency of an indicator in measuring the inequalities across the regions having different population size.

6. Theil and Gini indexes should be defined (as a mathematical formula) on page 7.

R: Thank you for your suggestion. The mathematical formulae for the relevant models have now been included accordingly.

7. Statements like "The GI is one of the most commonly used measures of distributional inequalities in healthcare …" should be either supported (by direct evidence or citation of literature sources) or skipped.

R: Thank you for the suggestion. References have been included to the revised manuscript now.

8. What do you mean by "This analysis utilises the value of observations greater than zero." on page 8? Does it mean that you excluded 0 values from modeling? If yes, how much it smooths (decreases the inequality of) the results?

R: We did not exclude zero observations from the analysis. However, as it was described in the results section, we used different Official Stata Commands for measuring the inequalities. The indices provide different values when there is zero value in the time-series observation(s). As in the case of the Theil L and Theil T indices in measuring the inequalities, we have commands which accommodate zero values and which exclude zero values from the analysis of the Gini inequality. The *"ineqdeco"* stata command provides measured values of the GI and most of the Generalised Entropy (GE) measures including Theil L and Theil T indices. However, the Gini value generated by this command is underestimated because it excludes the zero value from the analysis if it exists in the observation. Instead, the *"ineqdec0"* provides an appropriate measure of the Gini inequalities by accommodating the zero values in the dataset. Thus, we used different stata commands as appropriate to accommodate zero values in the time-series observations in measuring the Gini inequalities. However, the *"ginidesc"* official stata command for carrying out the Pyatt's Gini decomposition into three additive inequality components and into the non-additive sub-group (regions in our case) Gini values simultaneously, excludes zero values from the analysis.

9. Bootstrap and jackknife techniques mentioned on page 8 did not cope with the (potentially serious) effect of missing values, right?

R: We used the bootstrap and jackknife techniques as appropriate to determine the 95%CI for the overall Theil L, Theil T and Gini indices. Besides, we used the bootstrap technique to generate the 95%CI for the inequality changes at two points in time using the *"dsginideco"*, for Theil L, Theil T using *"ineqdeco"*, and for the relative marginal changes in inequalities using the *"descogini"* Official Stata Commands. These techniques were helpful for the stated purposes, not for handling the effects of the missing observation. For this reason, we estimated the missing observations using the annual growth rate of the distributions.

10. How exactly you fitted the multiplicative model on page 8? Did you take log of both sides, or did you even estimate different components on the right hand side separately? The statistical modeling and estimation procedures have to be described much more carefully and in substantially greater detail. This should include mathematical form of the model, assumption about error, concrete covariates and how they entered the model, estimation techniques used for identifying parameters etc.

R: We did not take log in the modelling. This multidimensional Gini inequality decomposition was performed using a single Official Stata Command *"descogini" available in the stata statistical software*. This command generates the values for the S_k (the GHE share of the explanatory variable), G_k (the Gini of the explanatory variable), R_k (the Gini correlation of the ranked explanatory variable with the overall GHE inequality) I_k (Share of GHE Inequality) presented in column 1 to 4 of Table 5 and the bootstrap technique for generating the "relative marginal changes" presented in column 6 of the same Table. The values in column 5 and 7 were calculated using the formulae for calculating the "Relative GHE inequality" and "Elasticity" of the explanatory variables. The formulae are presented in column 5 and 7. The appropriateness of the calculations is confirmed by the similarity of the "Relative GHE inequality" and "Elasticity" values as presented in Table 5 of our manuscript. The details of the techniques are provided by the authors in the following sources:

- *Lerman RI, Yitzhaki S. Income inequality effects by income source: a new approach and applications to the United States. The review of economics and statistics. 1985 Feb 1:151-6.*
- *Garner TI. Consumer expenditures and inequality: an analysis based on decomposition of the Gini coefficient. The Review of Economics and Statistics. 1993 Feb 1:134-8.*

11. There are too many unnecessarily detailed tables that are hard to read. Could you present some of the results in the much-more-easily-comprehensible graphical form? The temporal trends (as overlay of curves for different regions etc.) are particularly suitable for presentation …

R: As partly stated for reviewer 1, our objective was to analyse the aggregate level inequalities of the distributions. We used the *"ginidesc"* Official Stata Command for the Pyatt's Gini inequality decomposition which provides the overall inequality of a distribution, its additive three Gini components (the net inter-region, intra-region, and interaction term Gini components) and the non-additive Gini values for the regions simultaneously. Thus, we preferred to summarise the outputs in a table than using several graphs for many of the indicators. The tabular presentation is more informative about the extent and statistical significance of the observed inequalities, which could not be identified from graphical presentation of the Gini values for the regions. Nevertheless, we have considered the inclusion of the graphs concerning trends of inequalities for the aggregate level.

11. The revised version of the paper should be substantially shortened and made much more compact.

R: Thank you for your suggestion.

Reviewer 4:

1. The description of the data is so vague that it is nearly impossible to understand what the authors did exactly. It appears to me from the description in p.5 lines 21-34 (line numbers refer to the numbers in the PDF) the authors have one data point per region per year. However, in page 26 (of 28), Gini is reported for each region. Therefore, the authors must have gotten some distributional information for each region, but it is unclear how it was obtained. I am also not sure whether the figures in this page were created from HHRIs or EDHS (a kind of information that should be included in the figure).

R: Thank you for the comment and providing us the chance to improve the report. As it was stated in the manuscript, we had annual data for all the regions over sixteen-year period for all the indicators (2000 to 2015) except the data for under-five mortality and infant mortality rates, which were five-yearly observations for each region, i.e., it is a panel data. To make it clearer, now we have included figures showing the aggregate level trends of the Gini inequalities overtime in the revised manuscript.

Regarding Fig.1 & 2 presented in our manuscript, descriptions were given just below the references. The findings illustrated in Fig. 2 were the outputs of our analysis.

2. Further, what is reported does not make too much sense to me. What is meant by "population ranked by share of "nurses + midwives"? What is top 10% in this context? Top 10% of what? How is Gini computed? This is reference to which year? There are so many obvious issues that need to be addressed

R: From the two most common approaches of inequality analysis, *the place (geographic)- based inequality analysis and population (person)-based inequality analysis*, we used the population-based

inequality analysis of the distributions. In our study, the combination of the nurses and midwives was used as a single indicator. Thus, the proportion of a given distribution (e.g., nurses and midwives together (NMWs)) was divided by the cumulative proportion of the population. Thus, the population's share of this distribution for each region has to be put in ascending order and ranked accordingly to analyse the inequality or Lorenz curve. As it is known, in a Lorenz curve *the x-axis shows the cumulative proportion of population ranked by a given distribution (e.g., population sharing* NMWs *in our case)* and *the y-axis indicates the cumulative proportion of a given distribution (e.g.,* NMWs*) for a given proportion of population.* This description was stated in our manuscript. We applied the *"Stacked bar charts"* to present the graphs. This was done using the official stata command *"pshare stack, plabels("bottom 50%" "mid 40%" "top 10%")"*. This command is flexible and provides both the graphic presentation of a distribution and the Gini values for the regions analysed. The interpretations were also presented in our manuscript accordingly.

3. The authors make a very naive cross-country comparison. The level of measured inequality depends on the unit of analysis. Without addressing this issue, cross-country comparision in the avaialbility of healthcare personnel would be meaningless

R: We agree with your comment. Our primary objective was to measure aggregate level inequalities using the regional data. This was described in methods section of our manuscript. Also, as it is partly stated under #4 for reviewer 3, the regions are autonomous states, as Ethiopia is ethnic-based federal system. Thus, the Pyatt's Gini decomposition technique was used to disaggregate the overall Gini value of a distribution into an additive three components of the Gini (inter-region Gini, intra-region Gini and the interaction term Gini). This approach also simultaneously provides the non-additive Gini value for the sub-groups (regions).

4. Finally, the writing of the paper also needs to be substantially improvoed. There are obvious problems in writing (e.g., p.6 of 28, line 25. "The ratios for were weighted by a fixed number of people for all the regions." The sentence is grammatically incorrect. Also, what is meant by "fixed nubmer of people

R: Thank you for your constructive comment. Now revision has been made and has been edited by a native English speaker. All the changes are highlighted in Yellow.

VERSION 2 – REVIEW

REVIEWER	Masatoshi Matsumoto Hiroshima University, Japan
REVIEW RETURNED	16-Nov-2018

GENERAL COMMENTS	Authors revised properly according to my comments.

REVIEWER	Marek Brabec Institute of Computer Science Czech Republic
REVIEW RETURNED	16-Nov-2018

GENERAL COMMENTS	The paper improved substantially since the previous submission and is now ready for publication.