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)	Relative importance of pre- and postnatal determinants of stunting: data mining approaches to the MINIMat cohort, Bangladesh
AUTHORS	Svefors, Pernilla; Sysoev, Oleg; Ekstrom, Eva- Charlotte; Persson, Lars Ake; Arifeen, Shams E.; Naved, Ruchira; Rahman, Anisur; Khan, Ashraful Islam; Selling, Katarina

VERSION 1 - REVIEW

REVIEWER	Dr Sajid Soofi Aga Khan University Pakistan
REVIEW RETURNED	21-Aug-2018

GENERAL COMMENTS	The paper is exceptionally well written on important public health topic of interest specially for resource limited countries as stunting rates are alarmingly high in south Asian countries. This is an excellent study which provides reliable information regarding determinants of stunting. Title Question posed is focused in terms of population studied and the study tried to detect an effect. Abstract Abstract precisely delivered what has been found. Introduction Introduction is very coherent, provided with recent and updated references. Methods Methodology is appropriate to reach conclusion. Results Results presented very well, illustration were clear. Discussion Discussion and conclusion are well balanced and adequately supported by the data presented but it extra ordinary lengthy. Authors provided extensive explanation of statistical analysis and method/model used. Discussion section needs revision to make it more precise and definitive clarity for general audience. Please also reduce overall length of section as lot of unnecessary information is provided here.

REVIEWER	Chan Yiong Huak Biostatistics Unit Yong Loo Lin School of Medicine National University of Singapore
REVIEW RETURNED	03-Oct-2018

GENERAL COMMENTS	Figure 2. The word randomised should be removed. 'Recruited' would be more appropriate. Figure 2. 4436-845 does not equals 3625. kindly elaborate Predictors were 'data-mined'; recommend that a calculator for stunted risk based on the predictors to be developed (this would be nice).

REVIEWER	Mihiretu Kebede Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH Bremen, Germany
REVIEW RETURNED	04-Feb-2019

GENERAL COMMENTS	Comments The authors applied machine learning algorithm to identify the most critical pre-and postnatal determinants of linear growth from 0 to 24 months and stunting at 2 years and to identify subgroups with different growth trajectories and levels of stunting at two years. Although applying random forest seems practical and sound approach, the feature selection process in the model building and the write up of the paper is concerning. The paper may benefit if the authors address the following concerns. Abstract: Conclude your results that directly answer your research questions. Then, you may need to add one powerful implication/outlook/recommendation of your study. Strength and limitations of the study(just below the abstract) Line 17-25 on page 3, the authors wrote the algorithm automatically discovers complex interactions between predictors and the outcome. How do you discover complex interaction by using the model for the question which attempts linear relationships (that is one of your research questions)? Deep learning method would have provided better results. Introduction: page 5, please cite references at the end of the first paragraph. Methods: The authors mentioned they have included more 309 variables to develop the models. However, they have only 2723 observations. Machine learning algorithms generally perform better with big data. If we take the rule of thumb of “1 variable for 20 observations”, the authors would require more than 6000 observations. Otherwise, they need to have mechanisms to select the important features using parameters reduction or other feature selection methods.

Some of the variables are not also good enough to measure what was intended to measure. Example: wealth was measured using the number of saris owned by a woman, diet was measured by solid and semi-solid foods given to infant from one to 12 months. That may be the reason why diet did not come out as an important variable in the model.

In the methods, section it was mentioned 309 variables were used to build the model. On another line, variables were randomly selected from the complete set of variables. How do we know the variables with high importance are not missed? What about selecting variables with applying feature selection algorithms using information gain values instead of random selection?

Outcome definition: The change in HAZ score was measured by subtracting HAZ at 24 months from HAZ at birth. This should have been reversed to make the interpretation easier. Positive changes would indicate desirable HAZ changes and negative HAZ would indicate undesirable changes.

On page 10, line 26, the authors mentioned: they have used random forest to impute the data as simulation study shows this model provides accurate results. Is this from a previous study? If yes, cite. Or is it after you have tried two methods and found out that random forest was better? If yes, please mention how you compared the two and in what metrics RF outperformed K-means.

Mention how the data was split into a test, train and perhaps validation (??) Is there a possibility that a sample an observation would be selected for test and train set? If not, how do you avoid not to be double sampled? Perhaps, making your analysis codes public may make it easier.

Results: line 11 page 12, please change “fetal loss” to “fetal death”.

Line 42-43: mean-0.94 to mean=0.94

Table 1: What does a small gestational age mean? Was the number of saris used to measure wealth? Why not wealth index? Number of saris is just one variable.

Are the values for change HAZ score on Table 1 mean values? If so, please change it to “Mean ΔHAZ 0-24 months”.

Why was change HAZ score positive for preterm (190/2723)? Did height get reduced from what it was at birth?

Page 13: Please exclusively differentiate HAZ at birth and HAZ at 24 months. Then it will be easier for the reader which HAZ score the authors meant. I thought a derived attribute was mistakenly added in the model. It is not clear unless the reader sees the figures.

Line 25-34 page 14, again, the interpretation would have been easier if the subtractions were reversed and positive changes were considered as desirable gains of HAZ score.

Discussion: This section needs reorganizing. On the first paragraph, write the main results that directly answer your

	research questions. Then, discuss the main results in the subsequent paragraphs. Paragraph 2 and 3 are about strength and limitations, move them to the end of the discussion section. The authors mention, on last paragraph of page 15, the difference with the non-analysed groups did not no likely influence on the primary outcomes of the study. How did you confirm that? Delete the two paragraphs of the page 16 or move them to methods. The study did not have conclusions. You need conclusion as a separate section after the discussion.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Dr Sajid Soofi

Institution and Country: Aga Khan University, Pakistan

Please state any competing interests or state ‘None declared’: NO CONFLICT OF INTEREST

Please leave your comments for the authors below

The paper is exceptionally well written on important public health topic of interest specially for resource limited countries as stunting rates are alarmingly high in south Asian countries. This is an excellent study which provides reliable information regarding determinants of stunting.

Title Question posed is focused in terms of population studied and the study tried to detect an effect.

Abstract Abstract precisely delivered what has been found.

Introduction Introduction is very coherent, provided with recent and updated references.

Methods Methodology is appropriate to reach conclusion.

Results Results presented very well, illustration were clear.

Discussion Discussion and conclusion are well balanced and adequately supported by the data presented but it extra ordinary lengthy. Authors provided extensive explanation of statistical analysis and method/model used.

Discussion section needs revision to make it more precise and definitive clarity for general audience. Please also reduce overall length of section as lot of unnecessary information is provided here.

Response: Thank you for your comments on the paper. The discussion has been shortened.

Reviewer: 2

Reviewer Name: Chan Yiong Huak

Institution and Country: Biostatistics Unit, Yong Loo Lin School of Medicine, National University of Singapore

Please state any competing interests or state ‘None declared’: No

Please leave your comments for the authors below

Rev: Figure 2. The word randomised should be removed. 'Recruited' would be more appropriate.

Response: Thank you, this has been changed.

Rev: Figure 2. 4436-845 does not equals 3625. kindly elaborate

Response: This discrepancy is due to twins and triplets.

Rev: Predictors were 'data-mined'; recommend that a calculator for stunted risk based on the predictors to be developed (this would be nice).

Response: Thank you for the suggestion but we have no opportunity to include such a work in relation to this paper.

Reviewer: 3

Reviewer Name: Mihiretu Kebede

Institution and Country: Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH, Bremen, Germany

Please state any competing interests or state ‘None declared’: None declared

Please leave your comments for the authors below

Comments

The authors applied machine learning algorithm to identify the most critical pre-and postnatal determinants of linear growth from 0 to 24 months and stunting at 2 years and to identify subgroups with different growth trajectories and levels of stunting at two years. Although applying random forest seems practical and sound approach, the feature selection process in the model building and the write up of the paper is concerning. The paper may benefit if the authors address the following concerns.

Abstract: Conclude your results that directly answer your research questions. Then, you may need to add one powerful implication/outlook/recommendation of your study.

Strength and limitations of the study(just below the abstract)

Response: Thank you for the comment. The results that answer our research question are presented under the result section. Strengths and limitations are presented below the abstract.

Rev: Line 17-25 on page 3, the authors wrote the algorithm automatically discovers complex interactions between predictors and the outcome. How do you discover complex interaction by using the model for the question which attempts linear relationships (that is one of your research questions)? Deep learning method would have provided better results.

Response:

Decision trees do not model linear relationships but instead make predictions of the response for a given combination (interaction) of input variables, such as X_1<t_1 & X_2<t_2.This is why “this representation is popular among medical scientists”, see [HASTIE, p.305] for details.

Contrary to the suggestion made by the reviewer, deep learning models are problematic to use in public health and medicine because they are essentially black-box models: “in general, the difficulty of interpreting these models has limited their use in fields like medicine where interpretation of the model is very important” [HASTIE, p. 409]

Rev: Introduction: page 5, please cite references at the end of the first paragraph.

Response: Thank you, the reference is moved.

Rev: Methods: The authors mentioned they have included more 309 variables to develop the models. However, they have only 2723 observations. Machine learning algorithms generally perform better with big data. If we take the rule of thumb of “1 variable for 20 observations”, the authors would require more than 6000 observations. Otherwise, they need to have mechanisms to select the important features using parameters reduction or other feature selection methods.

Response:

We believe that the reviewer is referring to a rule of thumb that might be relevant to models in which no variable selection is done. In this case, decision trees do select variables automatically, picking up a relevant split variable at each tree-growing step, so this rule of thumb does not apply here. In particular, a variable having the “strongest association to Y” is selected at each tree-growing step in conditional inference trees; see [HOTHORN, page 655]. Note, that final decision trees in our manuscript (page 37 and 38) contain less than 10 variables out of these 309 originally available variables originally included and tested in the trees.

Rev: Some of the variables are not also good enough to measure what was intended to measure. Example: wealth was measured using the number of saris owned by a woman, diet was measured by solid and semi-solid foods given to infant from one to 12 months. That may be the reason why diet did not come out as an important variable in the model.

Response: In figure 1 all variables that were included in the analyses are stated. Several of these variables measured wealth and diet, including asset scores, breastfeeding, etc., but those variables were not selected in the process of creating the conditional inference trees.

Rev: In the methods, section it was mentioned 309 variables were used to build the model. On another line, variables were randomly selected from the complete set of variables. How do we know the variables with high importance are not missed? What about selecting variables with applying feature selection algorithms using information gain values instead of random selection?

Response:

Random selection of variables is a standard technique used when building random forests. The random forests analyses were created based on 3000 trees. “The idea in random forests is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables”[HASTIE, p. 588]. Despite performing such a random selection, the importance of variables can be nicely computed with random forests, see the details in [HASTIE, chapter 15].

Rev: Outcome definition: The change in HAZ score was measured by subtracting HAZ at 24 months from HAZ at birth. This should have been reversed to make the interpretation easier. Positive changes would indicate desirable HAZ changes and negative HAZ would indicate undesirable changes.

Response: The change in HAZ score was calculated by taking HAZ at 24 months – HAZ at birth. Thus a desirable change in HAZ becomes positive and growth failure becomes negative.

Rev: On page 10, line 26, the authors mentioned: they have used random forest to impute the data as simulation study shows this model provides accurate results. Is this from a previous study? If yes, cite. Or is it after you have tried two methods and found out that random forest was better? If yes, please mention how you compared the two and in what metrics RF outperformed K-means.

Response:

The comparative analysis of the RF-imputation and the K-NN imputation is provided in the Supplementation appendix; we refer readers to this appendix on page 10, line 24.

Rev: Mention how the data was split into a test, train and perhaps validation (??) Is there a possibility that a sample an observation would be selected for test and train set? If not, how do you avoid not to be double sampled? Perhaps, making your analysis codes public may make it easier.

Response:

The reviewer refers to the holdout principle that assumes dividing the data into training and test data. However, this method is known to be less efficient than the cross-validation method that we use in the manuscript because the holdout principle needs to sacrifice some portion of data for computing the test error: “Ideally, if we had enough data we could set aside a validation set and assess the performance of our prediction model. Since the data are often scarce, this is usually not possible. To finesse the problem, K-fold cross-validation use part of available data to fit the model, and a different part to test it.”[HASTIE, page 241].

The cross-validation method “directly estimates the average generalization error when the method is applied to an independent test sample”[HASTIE, page 241].

Rev: Results: line 11 page 12, please change “fetal loss” to “fetal death”.

Response: We would like to keep the term fetal loss as it is a well-known term.

Rev: Line 42-43: mean-0.94 to mean=0.94

Response: The mean HAZ at birth was minus 0.94, i.e. -0.94

Rev: Table 1: What does a small gestational age mean? Was the number of saris used to measure wealth? Why not wealth index? Number of saris is just one variable.

Response: Small for gestational age is when a child is born with a weight that is low (compared to international references) according to its' gestational age [ALEXANDER]. Number of Saris mother owns was included in the table as it came out as an important predictor for Stunting at 2 years.

Rev: Are the values for change HAZ score on Table 1 mean values? If so, please change it to "Mean ΔHAZ 0-24 months".

Response: Thank you it has been added.

Rev: Why was change HAZ score positive for preterm (190/2723)? Did height get reduced from what it was at birth?

Response: Height was not adjusted for gestational age at birth thus, preterm children will have a low HAZ although they had an appropriate length for age. As they had a low HAZ at birth the change will be less negative or even positive compared to children that were born within an appropriate age.

Rev: Page 13: Please exclusively differentiate HAZ at birth and HAZ at 24 months. Then it will be easier for the reader which HAZ score the authors meant. I thought a derived attribute was mistakenly added in the model. It is not clear unless the reader sees the figures.

Response: Thank you for your comment, however we don't understand what sentences you are referring to, we have searched through the documented and haven't found anywhere where which age is not specified for HAZ.

Rev: Line 25-34 page 14, again, the interpretation would have been easier if the subtractions were reversed and positive changes were considered as desirable gains of HAZ score.

Response: As described above, and according to most common practice, the desirable changes=positive changes. Se figure 3 or [SVEFORS] for an illustration of the growth pattern during this time period.

Rev: Discussion: This section needs reorganizing. On the first paragraph, write the main results that directly answer your research questions. Then, discuss the main results in the subsequent paragraphs. Paragraph 2 and 3 are about strength and limitations, move them to the end of the discussion section.

Response: The journal author guidelines instruct us,, and we as authors, prefer to discuss strengths and limitations before the main results but it is off course a matter of taste.

Rev: The authors mention, on last paragraph of page 15, the difference with the non-analysed groups did not no likely influence on the primary outcomes of the study. How did you confirm that?

Response: This is not confirmed as we do not have this information from the non-analysed group. It is, however, very unlikely with an influence as the differences and lost to follow-up were small. This is discussed on line 360-365.

Rev: Delete the two paragraphs of the page 16 or move them to methods.

Response: We have reduced this section but believe it is important to discuss the strengths and limitations of the used methods.

Rev: The study did not have conclusions. You need conclusion as a separate section after the discussion.

Response: The conclusions are part of the last paragraph in the discussion section. If the editor prefers it to be placed under a separate sub-heading we will off course adhere to that.

]ALEXANDER] Alexander GR, Himes JH, Kaufman RB, Mor J, Kogan M. A United States National reference for Fetal Growth. Obstetrics and Gynecology. 2015; 87: 163–168.

[HASTIE] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. Second Edition. New York, NY, USA: Springer New York Inc..

[HOTHORN] Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional

inference framework. Journal of Computational and Graphical statistics, 15(3), 651-674.

[SVEFORS] Svefors, P. et al., 2016. Stunted at 10 Years. Linear Growth Trajectories and Stunting from Birth to Pre-Adolescence in a Rural Bangladeshi Cohort D. O. Carpenter, ed. PLOS ONE, 11(3), pp.e0149700–18.

VERSION 2 – REVIEW

REVIEWER	Mihiretu Kebede Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH
REVIEW RETURNED	15-Apr-2019

GENERAL COMMENTS	I thank the authors for trying to address some of my previous comments. However, not all of them were properly addressed. I cannot recommend publication of this manuscript, at least in the current form. Given that the authors' novel application of machine learning methods in public health, I preferred to give them one more opportunity to revise this manuscript.

	Conclusion section of the abstract: Please conclude only your results, not from reviews of the literature. The BMJ authors' guideline advises "Do not go beyond your results." Your current conclusion does not have a conclusion about your results. Perhaps, it has a recommendation statement. You need to remove the phrase which reads "together with findings from recent reviews". Page 8: rewrite the strengths and limitations section. Add the limitations and reduce the strengths to balance. Currently, you wrote 4 strengths and 1 limitation. The first bulleted point is not needed. In addition, please structure this section to first write the strengths and then the limitations. Another main limitation is the study did not offer stratified analysis for sex. WHO recommends stratified analysis. Introduction: Add reference at the end of paragraph one. Methods: In my previous comment I suggested to reverse the calculation of the mean change HAZ as ΔHAZ = HAZat 24 - HAZ at birth. Please consider this comment or provide an explanation. The methods should include a description of how household asset score was calculated. Results: first paragraph: Either remove the phrase which reads "data not shown" or show the data in a supplementary material or cite a publication which suggests those results. I recommend changing the word "foetal loss" to "foetal death" to be consistent with the medical literature and ICD coding. The term "foetal loss" is more important to parents than medical literature. Others such as foetal demise, foetal loss, stillborn, or stillbirth are less commonly used terms. Foetal death is the standard terminology. Line 576-577, it reads (-2.04), while in the figure it was -2.0. Check Figure 7 and make it consistent. The conditional inference tree displayed in figure 7 is not described in the results. Describe in few lines! The composite variable "household asset score", and other individual variables such as number of cows, TV ownership, shalvar kamiz, presence of electricity, mattress, number of pairs of shoes the mother owns, saris, etc were added to the model. All the individual variables are used to calculate "household asset score". In machine learning, it is not recommended to add both the composite and the individual variables together in the model because the individual variables are correlated to the composite variable or derived attribute (in your case household asset score) corresponding to their relative weights. The authors themselves partly conceded the impact and stated it as a limitation (on paragraph 6) of their modelling strategy. Instead, I recommend removing correlated variables from the model. Move the sentence on line 573-574 to the end of the paragraph. Otherwise, it is confusing with the current order. Discussion: In my previous comment, I suggested restructuring this section. The authors failed to change or respond to my comment satisfactorily. The current form makes the discussion less interesting to follow as a reader. The paragraphs (2, 3 & 6) describing the strengths and limitations should be moved to the end of the discussion just before the conclusion paragraph (unfortunately, this manuscript does not have a conclusion). Does the BMJ open author's guideline suggest the strength and

	limitation of original research articles should be written right after describing the main findings? No, it doesn't. I still wonder why the classifier variables in Figure 6 & 7 are different from the most important variables ranked based on relative importance and displayed in Figure 4 & 5. Those variables classified as important variables (with high MSE) should have been the classifier variables on the decision trees. Example: mother's education, gestational age, mother's weight, asset score, mother's education, chest circumference, were much more important than father's education using the relative importance. It is puzzling father's education preceded these variables t in the conditional inference tree presented in Figure 6. The same is true for variables important variables in figure 5 and the decision tree in figure 7. This difference should either be explained or the analysis should be revisited. One of the most important goals of machine learning is prediction. This should be highlighted in the discussion section. I invite the authors to reflect their remarks in few lines regarding the potential benefit of using machine learning methods for public health research and its advantages over traditional statistical modelling. More worryingly, this research does not have a single conclusion statement. My last words, please conclude your results! You may not need a separate section for it. But, the last paragraph of your discussion should be the conclusion. Minor comments Title: Please change the semicolon to a colon. Please change "mean -0.94" to mean = -0.94

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 3

Reviewer Name: Mihiretu Kebede

Institution and Country: Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH

Please state any competing interests or state 'None declared': None declared.

Please leave your comments for the authors below

I thank the authors for trying to address some of my previous comments. However, not all of them were properly addressed. I cannot recommend publication of this manuscript, at least in the current form. Given that the authors' novel application of machine learning methods in public health, I preferred to give them one more opportunity to revise this manuscript.

- Conclusion section of the abstract: Please conclude only your results, not from reviews of the literature. The BMJ authors' guideline advises "Do not go beyond your results."Your current conclusion does not have a conclusion about your results. Perhaps, it has a recommendation statement. You need to remove the phrase which reads "together with findings from recent reviews".

Response: Thank you for your comment, the phrase has been removed.

- Page 8: rewrite the strengths and limitations section. Add the limitations and reduce the strengths to balance. Currently, you wrote 4 strengths and 1 limitation. The first bulleted point is not needed. In addition, please structure this section to first write the strengths and then the limitations.

Another main limitation is the study did not offer stratified analysis for sex. WHO recommends stratified analysis.

Response: Thank you. The section (line 42-50) has been changed to the following:

- Includes high-quality longitudinal data with low rates of missing data on child growth and a wide range of pre and postnatal household, family, and environmental factors, child characteristics at birth, infant feeding, and morbidity.
- Employs decision-tree-based methods that permit the inclusion of a high number of predictor variables, variables of different types and automatically discover complex interactions between predictor variables and include them in the model.
- Some potentially important determinants of linear growth were not present in the database.
- The study does not include stratified analyses for girls and boys

- Introduction: Add reference at the end of paragraph one.

Response: Thank you. We have added the following reference: Bayer R, Galea S. Public Health in the Precision-Medicine Era. N Engl J Med. 2015 Aug 6;373(6):499–501.

- Methods: In my previous comment I suggested to reverse the calculation of the mean change HAZ as ΔHAZ = HAZat 24 - HAZ at birth. Please consider this comment or provide an explanation.

Response: Please note that the mean change in HAZ is, and has previously been, calculated the way you suggest, i.e. ΔHAZ = HAZ at 24 months - HAZ at birth. As you have stated before, in this way a desirable change in HAZ becomes positive and growth failure becomes negative. The majority of the children analysed in this manuscript experience growth failure from birth to 24 months so that the change in HAZ of the children with the most desirable growth is still negative (not positive), although, less negative than for the children that had the most pronounced growth failure. In short, many children have negative delta-values on HAZ even if the delta variable has been set up the way you suggest. The growth pattern of the children is illustrated in figure 3. We have added this clarification on page 9 line 228-229.

- The methods should include a description of how household asset score was calculated.

Response: This is described on line 161-164 “Socioeconomic status was assessed based on a range of household assets, and a continuous household asset score, with a mean value of zero, was constructed based on a principal component analysis“ Pradhan M, 2003 [1].

- Results: first paragraph: Either remove the phrase which reads “data not shown” or show the data in a supplementary material or cite a publication which suggests those results.

Response: Thank you for your comment, the phrase has been removed.

- I recommend changing the word “foetal loss” to “foetal death” to be consistent with the medical literature and ICD coding. The term “foetal loss” is more important to parents than medical literature. Others such as foetal demise, foetal loss, stillborn, or stillbirth are less commonly used terms. Foetal death is the standard terminology.

Response: Thank you for your comment, fetal loss has been changed to fetal death.

- Line 576-577, it reads (-2.04), while in the figure it was -2.0. Check Figure 7 and make it consistent.

Response: Thank you for noticing this. It has now been corrected in Figure 7.

- The conditional inference tree displayed in figure 7 is not described in the results. Describe in few lines!

Response: The figure is described in line 352-355 and 359-364. References to the figures have been added to make it clear in the following way: “ The difference in Δ HAZ between the identified subgroups of children with the most negative change and the subgroup with the most positive change was 2·22 HAZ. Children who already had a low HAZ at birth (≤-2·33) had the most positive change in HAZ from birth up to 24 months (+0·18 HAZ), while children who were born with a HAZ above 0.19 had the most negative Δ HAZ (-2·04 HAZ) (Figure 7).”

- The composite variable “household asset score”, and other individual variables such as number of cows, TV ownership, shalvar kamiz, presence of electricity, mattress, number of pairs of shoes the mother owns, saris, etc were added to the model. All the individual variables are used to calculate “household asset score”. In machine learning, it is not recommended to add both the composite and the individual variables together in the model because the individual variables are correlated to the composite variable or derived attribute (in your case household asset score) corresponding to their relative weights. The authors themselves partly conceded the impact and stated it as a limitation (on paragraph 6) of their modelling strategy. Instead, I recommend removing correlated variables from the model.

Response: We thank the Reviewer for raising this interesting question. In the text, we give an explanation of why both the composite variable and the individual variables were kept, see page 21 line 621-629.

“Some of the included variables like “household asset score” are composite variables, which depend on individual variables like TV ownership, number of cows, etc. Presence of both composite and individual variables creates computational problems for traditional models like linear regression and for some machine learning models due to possible high correlation between the individual and the composite variables. However, CIT methods perform automatic variable selection by choosing the most relevant variable (with the strongest association to the response) at each decision tree split step [HOTHORN, page 655]. . Accordingly, these methods automatically choose either a composite variable or an individual variable at each split step based on the relevance of this variable to the response.”

- Move the sentence on line 573-574 to the end of the paragraph. Otherwise, it is confusing with the current order.

Response: Thank you for your comment but we cannot find line 573-574 in any version of the paper.

- Discussion: In my previous comment, I suggested restructuring this section. The authors failed to change or respond to my comment satisfactorily. The current form makes the discussion less interesting to follow as a reader. The paragraphs (2, 3 & 6) describing the strengths and limitations should be moved to the end of the discussion just before the conclusion paragraph (unfortunately, this manuscript does not have a conclusion). Does the BMJ open author’s guideline suggest the strength and limitation of original research articles should be written right after describing the main findings? No, it doesn’t.

Response: We have restructured the discussion accordingly.

- I still wonder why the classifier variables in Figure 6 & 7 are different from the most important variables ranked based on relative importance and displayed in Figure 4 & 5. Those variables classified as important variables (with high MSE) should have been the classifier variables on the decision trees. Example: mother's education, gestational age, mother's weight, asset score, mother's education, chest circumference, were much more important than father's education using the relative importance. It is puzzling father's education preceded these variables t in the conditional inference tree presented in Figure 6. The same is true for variables important variables in figure 5 and the decision tree in figure 7. This difference should either be explained or the analysis should be revisited.

Response: We agree that the phenomenon mentioned by the Reviewer can be puzzling. We provide an explanation in page 22 line 640-651.

"It can be noted that the CRF and the CIT models are not fully comparable. This can be explained by two factors. Firstly, many predictors that were important in the CRF model are relatively highly correlated and thus have a similar relationship to the response. Once one of these variables is selected by the decision tree in a split, there is a high chance that the remaining correlated variables (although also important according to the CRF) will not be picked up as the next splitting variable. Secondly, the CRF models and the CIT models cannot be matched directly. The CRF is a combination of many trees and is thus a more flexible model than a CIT. However, CRFs are nearly black-box models: the only interpretable information that these models deliver is the variable importance measure. On the contrary, CITs are "transparent" and interpretable models but have a smaller predictive power. This is another reason of why these models are not generally capable of efficiently embedding all the variables that are important in the CRFs. "

- One of the most important goals of machine learning is prediction. This should be highlighted in the discussion section. I invite the authors to reflect their remarks in few lines regarding the potential benefit of using machine learning methods for public health research and its advantages over traditional statistical modelling.

Response: Thank you for this suggestion, we have expanded upon this in line 630-639.

"Traditional methods like linear regression often have lower predictive power than data mining methods. In some cases, the traditional methods are not even possible to compute due to a high number of predictor variables and complex interactions. The method used in this work, Conditional Inference Trees, belongs to the class of Interpretable Machine Learning models and display precise information on the priority, size, and direction of the association of the predictors with the outcome. In addition, the risk group identification, including the prioritization and relevant cut-offs of risk factors, can be of high public health relevance for the design and targeting of appropriate interventions with the most significant benefit. Thus, we believe that the CIT framework has a large potential in public health and medical applications.

"

- More worryingly, this research does not have a single conclusion statement. My last words, please conclude your results! You may not need a separate section for it. But, the last paragraph of your discussion should be the conclusion.

Response: Thank you for the suggestion, a conclusion section has been added (line 666-673).

- Minor comments. Title: Please change the semicolon to a colon.

Please change “mean -0.94” to mean = -0.94

Response: Thank you for the suggestions, the title and sentence (line 320) have been changed accordingly.

1 Pradhan M, Sahn DE, Younger SD. Decomposing world health inequality. Journal of Health Economics 2003;22:271–93. doi:10.1016/S0167-6296(02)00123-6

VERSION 3 - REVIEW

REVIEWER	Mihiretu Kebede Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH, Bremen, Germany
REVIEW RETURNED	11-Jun-2019

GENERAL COMMENTS	I congratulate the authors for providing satisfactorily explanations in their response to my comments and greatly improving this manuscript. This manuscript is now suitable for publication.