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)	Impact evaluation of the free maternal healthcare policy on the risk of neonatal and infant deaths in four sub-Saharan Africa countries: A quasi-experimental design with Propensity Score Kernel Matching and Difference in Differences Analysis
AUTHORS	Dwomoh, Duah; Agyabeng, Kofi; Agbeshie, Kwame; Incoom, Gabriel; Nortey, Priscilla; Yawson, Alfred; Bosomprah, Samuel

VERSION 1 – REVIEW

REVIEWER	Thomas G Weiser Stanford University, USA
REVIEW RETURNED	23-Aug-2019

GENERAL COMMENTS	The authors report a significant improvement in neonatal mortality in the two countries implementing user fee waivers, with a 45% reduction base on a DiD analysis. However, overall mortality rates appeared to be dropping in all countries (based on my reading of table 1). The matching and controlling appear to account for some of the biologically explainable cause for childhood deaths, but obviously not all, and I was not able to see what kinds of social and financial improvements were also captured or considered in the modelling work (the authors note controlling for “baseline country characteristics” – p 13 line 9 – but do not explain what those are). I will be the first to admit that I am not a statistician, and while I am familiar with a number of the statistical techniques used here, I cannot comment on their appropriateness to these data or findings. Empirically, removing user fees has been show to improve access to care, and if access is the only barrier to reducing mortality then these are powerful findings. However, removing user fees and allowing populations to access poor quality, underfunded, and insufficiently resourced health systems would not be expected to improve outcomes, and I cannot really say from this work that improving access alone has generated these reported results (the authors clearly note they are not making a causal argument). 1. Table 1 reports neonatal and infant deaths in the 5 years preceding the survey for the years around time of user fee waiver implementation. Deaths in the no FMHP countries were significantly higher than deaths in the FMHP countries, indicating that something beyond user fees is driving this. Assuming this, why should user fees be implicated in the findings? 2. While the matching is powerful, even with the matching the results seem to indicate a fairly weak effect on mortality by eliminating user fees (based on my reading of table 3). And this is only demonstrated in a few of the models presented. To my mind this represents unaccounted effects that are also playing a role in driving down

	mortality. Any other explanations for this? 3. The elimination of user fees is important in achieving UHC, however revenue lost to the facility by eliminating these costs are, in my experience, typically not adequately made up by government or ministry subsidies to the facilities and providers. Thus providers and facilities reduce their use of consumables and force other forms of out of pocket payments (grey market purchasing of medications and consumables, provider service fees, or other "tipping" for services). How is this accounted for, and what is the experience of users in the settings where user fees have been "eliminated"? The work and the modelling demonstrate very powerful results of eliminating user fees for pregnant women seeking pre- and antenatal care, and I think that despite my critiques this would be worthwhile pursuing for publication. I recommend review by a statistician if this is not already being undertaken.

REVIEWER	Romain Pirracchio UCSF San Francisco, CA USA
REVIEW RETURNED	07-Dec-2019

GENERAL COMMENTS	I read with a lot of interest the manuscript bmjopen-2019-033356 with special focus on the statistical analysis. The authors have to be congratulated for getting access to a large amount of survey data from several country and for the fact that they attempted to overcome the potential biaises related to the cross-sectional and the observational nature by applying advanced statistical analysis. I will focus my review on what I believe is the most important part of the statistical analysis, i.e., the causal estimation of the impact of the policy on perinatal mortality. While I appreciate the use of a causal estimator, I found some limitations in the way the authors have conducted their analysis : - first and foremost, in the context of purely observational data, it is important to clearly state : i) what is the causal quantity to be estimated (the average treatment effet, a.k.a ATE ? the average treatment effect in the treated, ATT, etc.); ii) the assumed structure for the data generation mechanism (exposure, covariate, outcome and more importantly identify the relationship between the covariate, the exposure and the outcome) and iii) the set of (reasonable) assumptions needed to conclude causally. I am under the impression that this first very important step of a causal analysis was missing here. For example there is no description of the covariates deemed to be associated with the outcome, with the treatment of with both. This is crucial to identify the variables to be included in the PS model - There is no information regarding the PS mode : which variables, how were they selected, which modeling strategy (logistic regression or machine learning) - they is no information regarding the matching procedure (type of matching, etc.) - It seems that the authors did not use a variance estimator accounting for the matching procedure such as the Abadie-Imbens estimator for matching estimators - Multiple sensitivity analyses are provided in the tables but they is limited details in the method section to guide the reader through these analyses. For example, it is my understanding that alternative PS estimators are used, such as IPTW ? If this is the case, did the

	authors use standard IPTW or ATT-IPTW ? This is crucial since PS matching and standard IPTW do not target the same causal quantity - When interpreting the results, the authors seem to consider that the PS matching analysis can be used to conclude at the population level. This is not accurate since PS matching estimates the ATT, i.e., the exposure effect in the treated that could have been matched to an untreated. Connected to this, a very important piece of information is missing in the result section: the sample size before and after matching. I would also provide details about the population that was matched and the one that could not be matched since the results do not apply to the latter

VERSION 1 – AUTHOR RESPONSE

Reviewer # 1
Comment #1
The authors report a significant improvement in neonatal mortality in the two countries implementing user fee waivers, with a 45% reduction base on a DiD analysis. However, overall mortality rates appeared to be dropping in all countries (based on my reading of table 1). The matching and controlling appear to account for some of the biologically explainable cause for childhood deaths, but obviously not all, and I was not able to see what kinds of social and financial improvements were also captured or considered in the modelling work (the authors note controlling for "baseline country characteristics" – p 13 line 9 – but do not explain what those are).
Response #1
We agree with the reviewer that not all the biologically explainable causes for childhood deaths will be eliminated and we clearly indicated that as part of the limitations of the study. We indicated that although robust statistical modelling techniques were employed, people should not infer causality as several other factors including the social and financial improvements that reviewer mentioned which could influence child survival within these countries. The use of the secondary data has limitations as some obvious confounders could not be controlled since they were not captured in the data set. We clearly admitted that as part of the limitation. We have also corrected the statement on the baseline country characteristics. What we meant was that the difference in difference with the time fixed model automatically eliminates the effect of time-invariants variables at the individual and country level that could potentially influence child survival. It now now reads as "The fixed-effects model controls for all time-invariant differences between the individuals and the country level factors such as the differences in geographic location, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics".

Commen #2
However, removing user fees and allowing populations to access poor quality, underfunded, and insufficiently resourced health systems would not be expected to improve outcomes, and I cannot really say from this work that improving access alone has generated these reported results (the authors clearly note they are not making a causal argument).

Response #2
We do agree with the reviewer that is not only access that will influence negatively on child survival, all the other factors mentioned by the reviewer are equally important but the first step towards care for any government in our humble opinion is to provide access and gradually improve on the quality of care. Our argument was that these goverments in low and middle income countries should first provide access an th rest should follow. Without access there will be no quality of care. Once again we have already addressed the concern of the reviewer as part of the limitation.

Comment #3
Table 1 reports neonatal and infant deaths in the 5 years preceding the survey for the years around time of user fee waiver implementation. Deaths in the no FMHP countries were significantly higher than deaths in the FMHP countries, indicating that something beyond user fees is driving this. Assuming this, why should user fees be implicated in the findings?
Response #3
We humbly disagree with the reviewer on this. In fact that was the main reason why we tested for the parallel trend assumption which do not just look at the figures at one time point but whether there was a statistical significant difference in changes over time between the intervention and non-intervention group prior to the actual programme implementation. In our case we looked at the trend between 2000/2003-2007/2008 and the results as presented in Table 2 did not show a significant difference over time.
Comment # 4
While the matching is powerful, even with the matching the results seem to indicate a fairly weak effect on mortality by eliminating user fees (based on my reading of table 3). And this is only demonstrated in a few of the models presented. To my mind this represents unaccounted effects that are also playing a role in driving down mortality. Any other explanations for this?
Response #4
We humbly disagree with the reviewer on this. The effect is actually not weak as suggested. The effect is actually the interaction between Time and FMHCP indicated as (Time*FMHCP). Especially when the data is not coming from a randomized design, other unexplained factors may contribute to the observed change and in fact that is the main reason why we indicated that the inference from the model should not be interpreted as causal.

Comment # 5
The elimination of user fees is important in achieving UHC, however revenue lost to the facility by eliminating these costs are, in my experience, typically not adequately made up by government or ministry subsidies to the facilities and providers. Thus providers and facilities reduce their use of consumables and force other forms of out of pocket payments (grey market purchasing of medications and consumables, provider service fees, or other “tipping” for services). How is this accounted for, and what is the experience of users in the settings where user fees have been “eliminated”?
Response # 5
We perfectly agree with the reviewer that revenue lost to the facility by eliminating these costs are typically not adequately made up by government or ministry subsidies to the facilities and providers. Even in Ghana which I can personally attest to that, there are still challenges with timing of payment of premium to these health facilities. However, our study did not cover experiences of users and health facilities in terms of subsidies that they receive from Government or ministries from services rendered. It was beyond the scope of the current study. We only sought to determine whether the policy has contributed to reduction in neonatal and infant deaths.

Reviewer # 2
Comment #1
-first and foremost, in the context of purely observational data, it is important to clearly state : i) what is the causal quantity to be estimated (the average treatment effet, a.k.a ATE ? the average treatment effect in the treated, ATT, etc.); ii) the assumed structure for the data generation mechanism (exposure, covariate, outcome and more importantly identify the relationship between the covariate, the exposure and the outcome) and iii) the set of (reasonable) assumptions needed to conclude causally. I am under the impression that this first very important step of a causal analysis was missing here. For example there is no description of the covariates deemed to be associated with the

outcome, with the treatment of with both. This is crucial to identify the variables to be included in the PS model

Response # 1

We agree with the reviewer on the fact that we were not clear enough on the causal quantity to be estimated. We estimated average treatment effect (ATE) using propensity scores with Kernel weighting adjustment. But we also presented the results on ATT for comparison purposes for the reader. We have included that statement in the methods section and it reads as “We estimated average treatment effect (ATE) using propensity scores with Kernel weighting adjustment and inverse probability of treatment weighting (IPTW). We clearly stated the variables that were used in the PS model. Perhaps it was not very clear to the reviewer which we admit. We have revised the section heading under the methodology section to read as “Covariates assumed to be associated with child survival and included in the estimation of the propensity scores”

Comment #2

There is no information regarding the PS model: which variables, how were they selected, which modeling strategy (logistic regression or machine learning)

- there is no information regarding the matching procedure (type of matching, etc.)

Response # 2

We partly agree with the reviewer. The variable selection was based on the analystical frame from Mosley and Chen which we clearly indicated. The estimation of the propensity scores were based on the binary logistic regression model. We have now added that to the methods section.

This is a very important observation from the reviewer and we have clarified that in the whole document. Basically we used propensity scores with kernel matching (weighting adjustment). We have made the changes in the document.

Comment #3

It seems that the authors did not use a variance estimator accounting for the matching procedure such as the Abadie-Imbens estimator for matching estimators

Response

We humbly disagree with the reviewer on this. The data set used originate from a complex survey design. To obtain robust standard error for inference, the complex survey design structure (weighting, stratification and clustering) were all adjusted for in both the propensity score model and the final model using Taylor linearization robust sandwitch estimator.

Comment#3

Multiple sensitivity analyses are provided in the tables but they is limited details in the method section to guide the reader through these analyses. For example, it is my understanding that alternative PS estimators are used, such as IPTW ? If this is the case, did the authors use standard IPTW or ATT-IPTW ? This is crucial since PS matching and standard IPTW do not target the same causal quantity.

Response #3

We used both (standard IPTW or ATT-IPTW) but the focus was on the standard IPTW to estimate ATE. We thought the detail was not necessary but we agree with reviewer and we have added a more detail on the two approaches with the relevant formulae.

Comment#4

- When interpreting the results, the authors seem to consider that the PS matching analysis can be used to conclude at the population level. This is not accurate since PS matching estimates the ATT, i.e., the exposure effect in the treated that could have been matched to an untreated. Connected to this, a very important piece of information is missing in the result section: the sample size before and after matching. I would also provide details about the population that was matched and the one that could not be matched since the results do not apply to the latter.

Response #4

We agree with reviewer on his opinion but we clarify as follows: We applied propensity score with Kernel matching (weighting adjustment) and ATE-IPTW which basically rely on all the sample size instead of 1:1, 1:2 nearestneighbour maching and other matching techniques. We have already clarified this in the main document to reflect the fact the matching was based on Kernel weighting adjustment and ATE-IPTW

VERSION 2 – REVIEW

REVIEWER	Romain Pirracchio UCSF, USA
REVIEW RETURNED	04-Jan-2020

GENERAL COMMENTS	I would like to thank and congratulate the authors for their revision and their thorough answers to my first set of comments. Although I generally believe they appropriately answered them, I still have some concerns : - Kernel matching : the new version of the manuscript explains much more clearly the methods and especially kernel based matching. However, some confusion remains related to the fact that the authors use kernel matching or kernel weighing to designate the same analytical procedure. Although I understand Kernel PS matching relies on individual weighting, I would strongly advocate using one of the 2 terminology (Kernel matching or kernel weighting) to avoid any confusion - Table 3 : Although I appreciate the fact that the authors challenged their main analysis using multiple alternative modeling approaches, it is difficult to compare aRRs, to HRs or ORs... In addition, I am confused by the result provided for ATET weighting and PSM- IPTW : aRR = -3.41**[0.46, 0.81] ? - Sample size post matching: I am only partly satisfied by the answer provided by the authors. Although I understand that kernel matching maximizes the chance of matching a control to a treated, observations outside the range of common support are still discarded. Can the authors further comment on that ? - Variance estimation : I respectfully disagree with the answer provided by the authors regarding variance estimation. I cannot see how a "simple" robust sandwich variance estimator could take into account the multiple sources of variability related to : PS estimation, individual clustering and weighting, PS matching and treatment effect estimation. This point should be more clearly discussed.

VERSION 2 – AUTHOR RESPONSE

Reviewer's comment
Kernel matching: the new version of the manuscript explains much more clearly the methods and especially kernel based matching. However, some confusion remains related to the fact that the authors use kernel matching or kernel weighing to designate the same analytical procedure. Although I understand Kernel PS matching relies on individual weighting, I would strongly advocate using one of the 2 terminology (Kernel matching or kernel weighting) to avoid any confusion

Response
We agree with the reviewer. We have now used Kernel matching throughout the entire document.

Reviewer's comment
- Table 3: Although I appreciate the fact that the authors challenged their main analysis using multiple alternative modeling approaches, it is difficult to compare aRRs, to HRs or ORs... In addition, I am confused by the result provided for ATET weighting and PSM- IPTW : aRR = -3.41**[0.46, 0.81] ?

Response
We agree with the reviewer that you cannot compare aRRs, HRs or ORs and in fact were not comparing effect size estimate directly, but we wanted to demonstrate that the intervention had an impact even when different models were fitted and that the results is not dependent on the choice of the statistical model. The results 3.41**[0.46, 0.81] was entered in error. It was a typo. We have revised it to 0.61**[0.46, 0.81]

Reviewer's comment
- Sample size post matching: I am only partly satisfied by the answer provided by the authors. Although I understand that kernel matching maximizes the chance of matching a control to a treated, observations outside the range of common support are still discarded. Can the authors further comment on that?

Response
We agreed with the reviewer. We have now added a lack of common support of the use of Kernel Matching as part of the limitations of the study.

Reviewer's comment
- Variance estimation: I respectfully disagree with the answer provided by the authors regarding variance estimation. I cannot see how a "simple" robust sandwich variance estimator could take into account the multiple sources of variability related to: PS estimation, individual clustering and weighting, PS matching and treatment effect estimation. This point should be more clearly discussed.

Response
We appreciate the concern of the reviewer. What we meant was that in the estimation of the propensity scores, we adjusted for clustering, weighting, and stratification. We have now made this clear in the methods section. We also used the sampling weight times the Kernel weight in final outcome analysis. We have cited the reference for using this estimation technique in the methods section already.

VERSION 3 – REVIEW

REVIEWER	Romain Pirracchio University of California San Francisco USA
REVIEW RETURNED	18-Feb-2020

GENERAL COMMENTS	I commend the authors for their work. They adequately addressed my comments and I have no other concerns.