Response to comments for article submission titled "Using Mobile Phone Data for Epidemic Response in Low Resource Settings - A Case Study of COVID-19 in Malawi" (DAP-2020-0037) to Data & Policy

Dear Reviewers,
Thank you kindly for the thoughtful comments, questions, and insights which have been valuable to improving our work. Please find below a point-by-point response (in red) to each comment with reference to the appropriate line reference in the revised submission.
Please do revert if anything has not been sufficiently addressed or we have still misunderstood.

Reviewer: 1

Comments to the Author

This is a very interesting and well presented piece of work which exemplifies the potential to re-use routinely data , in this case mobile phone call records, to support public health interventions. The approach is particularly applicable to low and middle income settings where smartphone ownership and usage may be low.

My specific comments and suggestions on how the paper can be improved are listed below. As a general point there should be enough detail in the methodology to enable others to reproduce the methods and in some places I felt this was lacking , I have detailed those that I identified below.

The authors should carefully check that all statements of fact are properly referenced e.g. p2 line 44 (re risk factors) , p3 line 69 (re test positivity rates) and p 3 line 74 (smartphone apps) - this is not an exhaustive list
We have reviewed the aforementioned gaps in referencing, as well as a few other key areas and added the relevant reference.

I would suggest in line 52 reference is made to the new COVID-19 variant identified in SA and its potential to spread throughput the region .

>We have added a reference to the B.1.351 variant and its importance to the second wave seen in Africa, and Malawi on lines 55-57.

p3 line 87 , add reference to the year (2020)
>Added, now on line 89

pages 4 lines 89 – 91 - I would like to see you expand on this as it wasn’t clear how you used the Census data to validate the population size estimates from mobile phone subscribers
>We have elaborated on these methods by adding a section on validation of the population estimates on lines 146-152

A general point – may need to go in the limitations is related to the potential for differential use of phones i.e. some mobile users will make many more calls/texts than others and whether this could skew your interpretation of the data e.g. in line 129 you assume I think that all subscribers make equal use of their phones in order estimate pop size in a TA at a particular time.
>This point is well taken. We have added a limitation specifically on the fidelity of information obtained in CDR data relative to mobile phone use/frequency on lines 317-319.

Line 99 – It wasn’t clear if IRB approval had been given for this particular study or for the previous work . Please reference the review board name and the approval reference number. Given the potential for aspects of this data to be misused to monitor the population or individuals for reasons other than health , I would like to see that there has been independent review of the research specifically carried out in this study.

>We have revised this portion and added the reference and protocol numbers for the NCST in Malawi, on lines 108-110.

Line 120 – Can you explain ‘salted-hashing’
>We have elaborated in explaining salted-hashing and its rational with supporting citations on lines 130-135.

Line 161 – Can you explain why each node was a cluster of cell towers and not just a single tower.
>We have clarified this point, which was incorrectly written before. In truth, only some nodes are representative of a cluster of towers, not all of them. Specifically this is the case in densely populated urban areas where multiple towers are assigned the same location in an effort to boost service availability and quality due to high numbers of users (line 193-197).

The paragraph from line 165 – 170 wasn’t clear , the methodology needs expanding to clarify exactly what was done . My interpretation is that you create a super-set of all individual journeys between nodes and use this to show movement patterns . The assumption is that infection spread follows movement. This is somewhat true for a respiratory infectious disease but would also be dependent on individual behaviour and the contact between individuals. I would like to see this discussed in more detail so I can comment in detail on the validity of the approach.
>Apologies for the limited detail in this section. We have revised considerably, expanding on the basis for the analysis as well as the technical approach. Lines 182-210 are intended to comprehensively explain why and how we leverage the CDR data for mathematical modeling, specifically to generate the epidemic start dates in TAs presented in Figure 5.

Line 198 – What is meant by (on a rolling basis) ?
>This has been removed, as we intended “daily” which was already captured in the statement, now line 222-223.

Line 198 – How did you remove marketing accounts ?
>These accounts are marked within the MNO’s records, and therefore were excluded from transfer before the analysis began (line 223-224).

Line 199 – is the 1.5 million an average over time ?
>Yes, we meant average per day, now revised on line 224.

Line 200 – You should justify your statement that the CDR data was highly correlated to population . Did you calculate a correlation coefficient ?
>We have added some detail on the validation of the CDR data to population estimates (line 227). Specifically we estimated the deviation from census and worldpop estimates at 5% or less across TAs.

Line 255 – 256 – I didn’t understand this sentence please expand to explain how you estimated epidemic start dates . Also Fig 5 has no key , title or access labels . If you are aiming to show the epidemic spread this would be much better as a tiled series of snapshots at for example monthly intervals. Is the colour representing number of infections ?
>Per the previous comment, we have expanded significantly in the methods section (line 182-210) on how this analysis was performed. We refer you to that section for the how, and have further added some detail on the results form lines 252-255. We have also updated the figure 5 and added a legend.

I would be interested to know (though accept that this may be beyond the scope of this paper) the extent to which this data has been used by the Malawian MoH to direct their response to the epidemic.
>This is a well received question. Reviewer 2 had a similar request, and we have added some details in the discussion on lines 269-273 as well as lines 293-297. Briefly, these analyses were underdevelopment throughout much of the first wave of COVID-19 in Malawi, and while partly insightful were difficult to leverage during the urgent emergency situation. However, during the second wave (which occurred while this manuscript was under review), the opportunity arose to leverage many of these tools for the response. Specifically the official mathematical model for the ministry of health which uses the projected epidemic start dates by TA.

Reviewer: 2

Comments to the Author

This paper makes a valuable contribution to our understanding of how CDR data can be used to model the potential spread of communicable disease and the effectiveness of policies aimed at restricting mobility. It provides useful detail on some of the technical systems that need to be in place to support such a project and on how the team addressed concerns related to privacy and potential bias stemming from the data used.

The one area where I would have liked to see more detail is on exactly how health authorities in Malawi ended up using this data and how much value they derived from it. Is the information from the CDRs now an indispensable input into their decision-making or is it less important? Answering these questions could make for a useful follow-on briefing.

>Reviewer 1 had a similar question, which we have added some details on lines 269-273 and 293-297. Briefly, these analyses were underdevelopment throughout much of the first wave of COVID-19 in Malawi, and while partly insightful were difficult to leverage during the urgent emergency situation. However, during the second wave (which occurred while this manuscript was under review), the opportunity arose to leverage many of these tools for the response. Specifically the official mathematical model for the ministry of health which uses the projected epidemic start dates by TA.

Less substantively, it seemed odd that the "competing interest statement," "funding statement," and "data availability statement" were floating in the middle of the paper rather than put at the end.
>These sections have been moved to the end of the paper on lines 339-356.