Author’s response to reviews

Title: A health systems strengthening intervention to improve quality of care for sick and small newborn infants: results from an evaluation in district hospitals in KwaZulu-Natal, South Africa.

Authors:

Christiane Horwood (horwoodc@ukzn.ac.za)
Lyn Haskins (haskins@ukzn.ac.za; lynh@saol.com)
Sifiso Phakathi (sifisophakathi@gmail.com)
Neil McKerrow (NEIL.McKERROW@kznhealth.gov.za)

Version: 2 Date: 05 Nov 2018

Author’s response to reviews:

To the editors BMC paediatrics:

Thank you very much for the helpful reviewers comments and the opportunity to respond to these. We have responded to each comment from the reviewers below and made the suggested changes. All changes are highlighted in the text. If there are any further queries do not hesitate to come back to us.

Reviewer 1:

Preamble

I appreciate the opportunity to review your work on "A health systems strengthening intervention to improve quality of care for sick and small newborn infants: results from an evaluation in district hospitals in KwaZulu-Natal, South Africa." This manuscript presents results from a pre-/post-study of 39 district hospitals indicating improvement in selected quality indicators (mostly inputs to care) over the three-year intervention period. As noted in the introduction, reducing preventable newborn morbidity and mortality is an important global health challenge and a priority area for health system interventions. The major strength of the work is the multifaceted intervention delivered at provincial scale; drawbacks include weak study design for drawing causal inference, incomplete assessment of quality, and insufficiently detailed methods.

The evaluation does not have a comparison group (all district hospitals in the province participated) or details on trends in hospital performance before or after the intervention period. The quality indicators chosen follow locally relevant standards, though focus on inputs to care
(infrastructure, equipment, provider knowledge) combined with a record review. Details of the intervention implementation are sparse and the calculation of some elements of the quality scores could be more clear; the methods require additional detail to enable comprehension and potential replication of this work.

Were these issues, detailed below, to be addressed, this article could provide useful initial evidence of intervention outputs, though it will not speak to causal effects or potential changes in health outcomes. As such, it provides a contribution to implementation science more than to new knowledge around health system performance for newborns; authors may consider expanding the implementation details to bolster this contribution to the literature. I believe the gaps in the methods can be fully addressed while the other considerations can be acknowledged and discussed more completely. Please see specific comments below addressing areas of concern.

Comment 1:
1. Study design

While the real-world nature of this intervention clearly precludes some randomized study designs, the pre-/post- designs provides relatively limited insight on the effect of the intervention as opposed to general trends or other health system strengthening. Even within the health worker training, availability of a pre-training test for all health workers would help to isolate the training itself, rather than selection of more capable health workers, as an explanation for the higher scores.

Authors response: We acknowledge the limitations of the design of the study in the real world setting that does not allow us to infer that changes are directly attributable to the intervention and we have strengthened the discussion of the limitations of the study to reflect this (L381-394). Similarly, we acknowledge that the selection of HWs for training was not clearly defined, so that we cannot confidently attribute improvements to the training. We have added a note to this effect in the discussion of limitations of the methodology. However, given that this was the only initiative directed at improving quality of newborn care during the study period, and the consistency of the trends in scores shown over a large number of hospitals, we suggest that the results can provide valuable information for others aiming to undertake similar work.

Comment 2:

Are there no external data at all that might be brought to bear, like comparison of accreditation in this province vs. a similar province without this intervention, routine reports of morbidity and mortality that start before the intervention and continued after it, etc.? Assuming not, the manuscript would be strengthened by reducing the claims of causality in the discussion for the
intervention as a whole and for the health worker knowledge, with clear consideration of alternative explanations for the results found and their relative plausibility in this setting.

Authors response: Unfortunately there is no information about accreditation in other settings or in places without any intervention to provide a comparison. In addition, the Department of Health (DoH) has sought to improve quality of care (QoC) for newborns using a number of strategies over the past decade. However, during the three-year implementation period of KINC, this was the only initiative used by the DoH to improve newborn QoC. However, we agree that in the real world setting, and without a control group, we cannot infer causality or attribute improvements to KINC interventions alone. We have added further note of this in the discussion (L388-393). In addition, we have also reduced the strength of the statements about causality in the discussion.

We have data on general trends in neonatal mortality from the routine district health information system (DHIS). DHIS data is unreliable and incomplete, and efforts have been made to improve this, so that more deaths are captured. As a result recent trends are difficult to interpret, but there is no evidence of any trend of improvement in neonatal mortality during and after KINC implementation (see DHIS data given below). However, given the complex aetiology of neonatal deaths and the poor quality of the data, we would not expect to see such a trend. We did not think it was possible to show this data in the manuscript, but we have added an additional note to say that routine data does not indicate any trend of improvement and added a reference to support this (L395-398)

<table>
<thead>
<tr>
<th>Year</th>
<th>Early NN death rate/1000 live births</th>
<th>NN death rate in facility/ 1000 live births</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>8.1</td>
<td>9.6</td>
</tr>
<tr>
<td>2010</td>
<td>8.8</td>
<td>10.2</td>
</tr>
<tr>
<td>2011</td>
<td>8.8</td>
<td>10.2</td>
</tr>
<tr>
<td>2012</td>
<td>9.2</td>
<td>10.9</td>
</tr>
<tr>
<td>2013</td>
<td>11.1</td>
<td>14.6</td>
</tr>
<tr>
<td>2014</td>
<td>9.8</td>
<td>12.4</td>
</tr>
<tr>
<td>2015</td>
<td>10.6</td>
<td>13.4</td>
</tr>
<tr>
<td>2016</td>
<td>10.2</td>
<td>12.9</td>
</tr>
<tr>
<td>2017</td>
<td>10.1</td>
<td>12.5</td>
</tr>
</tbody>
</table>
Comment 3:

For the latter it is not made clear how staff were selected for training and if these are even the same staff being compared - perhaps the staff not trained left and were replaced with newer staff, while those who were trained stayed and continued to gain experience, just as one alternative explanation.

Authors response: Staff were selected to participate in training by managers in each hospital. KINC training was conducted in partnership with the DoH, and routine DoH systems were used to invite participants. The only instruction was that staff selected to attend the training should be working in the neonatal nursery. We have added a note to this effect (L159-160). There is no indication that participants were not representative of the staff complement, particularly in small nurseries where there are few staff members to choose from. However, we cannot be sure of this and it is certainly possible that trained staff were systematically different from untrained, for example more experienced staff may have been selected, or those in leadership positions in the maternity unit. A further note has been added into limitations of the study that we cannot infer that improvements in knowledge were entirely due to KINC training (L382-385).

Comment 4:

And 2) considering more emphasis on the lessons learned from implementation if these are not reported in a separate paper. This would include challenges such as staff turnover, consideration of receptivity of the hospitals and their staff, adaptation of the intervention to context, differences between hospitals due to baseline differences in patient volume / resources / leadership, etc.

Authors responses: We have added additional points in the discussion about the challenges experienced in KINC implementation and how these were addressed (L305-312)

Comment 5:

More detailed reporting on these elements of the intervention could build on the strength of the study as a second contribution of the manuscript beyond the evaluation itself. There is a large body of work considering mentoring, supervision, and training as quality improvement interventions; understanding the attributes (type of training, duration, pedagogical methods, elements of supervision) applied in this case and how the results compare to other similar studies would strengthen the inference that could be drawn from this study.
Authors responses: We have considerably strengthened the description of the intervention, adding more detail about the processes followed during each activity (see under heading description of the intervention)

Comment 6:

More expansive discussion of accreditation early in the work would be helpful, as this is mentioned only briefly in methods and results but then discussed at length; again there are several existing studies assessing whether accreditation can achieve quality improvements (particularly given its cost) that could enrich the discussion.

Authors response: We have added a little more detail about accreditation, the accreditation process has been described in detail in a separate paper that it referred to in the text (ref given below) (L169-173)


Comment 7:

2. Quality assessment

Quality of care is measured here as primarily inputs to care, with some staff capacity and adherence to guidelines included. The authors acknowledge the limitations of these measures, but do not explain why outcomes such as severity of disease, length of stay, and mortality were not considered. Using outcomes as quality measures involves several considerations such as rarity of severe outcomes and importance of adjusting for baseline risk, but for newborns in particular most outcomes are health system sensitive and worthy of consideration. Because the volume of deliveries is not reported, it is not possible to say if adequate power for these outcomes might have been available. Please justify the decision not to consider these outcomes; if data are not routinely collected on them, should that not have been addressed as part of the intervention in order to enable ongoing monitoring? A more complete explanation is warranted.

Authors response: We have added a mention of the number of deliveries per annum in the largest and smallest district hospitals to highlight that district hospitals included a broad range from very large to extremely small hospitals (L123-125). This was a challenge and means that the hospitals varied widely in resources and competencies at baseline, as is shown in Figure 2. It would not be possible to include more detail on the number of deliveries in individual hospitals given the large number of participating hospitals.
The assessment tools were developed in collaboration with paediatricians from the DoH, and were reviewed and agreed by this team. The assessment of quality of care aimed to be as comprehensive as possible including as many items as possible that contributed to quality of care. However, these elements needed to be objective, clearly defined and consistently measured by a single field worker within a reasonable time (usually one day). We included as many variables as possible that could be directly observed, rather than relying on reported behaviours or attributes, which were open to bias. We have added a note to say that reporting bias is a limitation (L388-389).

If such a scoring system is to be used, it has to be valid and reliable, and usable by data collectors who are not experts in neonatal care. Thus, although we carefully considered assessing outcomes for admitted babies, after piloting we decided not to include these indicators. This was because of the complexity of assessing clinical care provided, the huge variety of presentations and conditions found in the neonatal nursery, and the multiple factors that contributed to outcomes. In small hospitals with few admissions, a single baby could change the overall assessment. It was difficult to compare the management of prematurity with management of jaundice, and assess outcomes in different babies with different complications, particularly when you consider the wide variety of facilities included within the umbrella of district hospitals, and the huge differences in the accessibility of referral services. Finally, we also considered that this data collection was by a non-paediatrician. We have added a note about the reasons for this decision (L180-182) and added a section on the limitations of the tool in the discussion (L360-366).

We therefore included record reviews, which focussed mainly on simple and universally applicable record keeping and observations undertaken. However, we acknowledge that some important elements of quality of care are neglected and we note in the discussion how the methods to be strengthened to improve this in future. (L371-377)

Comment 8:

3. Methods

The development of the quality scores, most notably the record review and observation of staff, requires additional detail to enable readers to understand it and researchers to consider replicating these methods. For instance, the methods section indicates that 10 points were assigned to records, but only 7 items are listed. It also indicates that 10 record reviews (selected at random?) were completed, but based on the numbers in results, it would be an average of 4 - 5 per hospital at each time point.

Authors response: the number of records included in the record review was five and the records were of the last five babies to be discharged from the neonatal nursery. The score for each record
was out of 10. We have clarified the explanation of the scoring system for the record review (L207-216) and how the record review was conducted (L191-192).

Comment 9:

The role of the observation of staff in calculating quality scores is not fully explained. Please include for both record reviews and staff an explanation of the desired sample size, the method of sampling, the final sample size per hospital, and the method of averaging within hospital to develop the quality indicators.

Authors response: The reasoning was similar to that described in response to comment 7 above. We recognised that skills were important in determining quality of care, but skills testing was limited to simple and important skills applicable in all neonatal nurseries and to all health workers. We also limited to those skills where correct performance could be clearly defined, and which could be easily and quickly assessed. We have added an additional note in the text describing the skills assessments and how these were conducted (L185-190).

Although the field worker was requested to randomly select different HWs on duty on the day of observation this was done on an ad hoc basis, and a structured process was not followed to select the participants for the skills assessment. We did not calculate a sample size but made the decision to include four skills assessments at each hospital. These are limitations to the methodology and we have added a note about this in the limitations of the study (L381-388).

Comment 10:

Please link the quality indicators to their source in each table. Table 2 listing care practice results includes structural indicators (KMC beds, bottles and teats not on view, health cards available) as well as individual-level items and facility-level items such as perinatal review meeting minutes. It is not clear if the staff practices are based on record review or observation of staff, and if so how many staff were observed per hospital and how the summary statistics were then calculated.

Authors response: We have added a column indicating the source of the data to each of the relevant tables as suggested. Wherever possible, direct observation was used to collect data and involvement of staff was minimised to avoid reporting bias, but this cannot be excluded, particularly for those variable that could not be directly observed. We have added a note on reporting bias under study limitations (L388-389). Staff practices were based on direct observation of the skills.

Comment 11:
Line 178-179 in the methods states that 'Scores were combined from the three domains to calculate an overall score for each hospital [...] all variables contributed equally to the final score.' Were the three domain scores averaged or were all items averaged without regard for domain? If the domain scores were averaged, then variables contributed equally to each domain and each domain contributed equally to the final score, but the weight of each variable may differ if the domains have different numbers of items, as I believe they do. Throughout the abstract and results, it would be helpful to present the scores including the total possible score to provide the scope of quality gaps.

Authors response: The scores were added together and divided by the total, rather than each score averaged and added to the total. Therefore, each variable contributed equally to the overall score. Dividing the items into domains was useful to clarify the process, but we did not feel that it was essential to give each domain an equal weight. A note has been added in the text to clarify this (L220-221). We have also added the total possible score to all scores presented wherever possible, as suggested by the reviewer.

Comment 12:

4. Statistical analysis

Only health care worker training scores are presented with p values to enable statistical inference, although the sample size is not provided and the statistical test applied is not described in the methods. Decisions on presenting statistical tests should be consistent throughout the work and described fully in the methods. While the sample size is small, it is not so small as to preclude appropriate testing such as Kruskal-Wallis or t tests of the quality scores.

Authors response: Thank you for the comment. For the knowledge score, the non-parametric Mann Whitney U test was used to test differences in the knowledge score for KINC trained and non-KINC trained participants. We have added a note on this in the analysis section (L222). Scoring for the knowledge scores are further explained in the text (L224-229)

The Wilcoxon Signed Rank test was used to test the difference between the hospital baseline score and the end points score. Again, we have added a note to this effect in the analysis section (L226). This showed that there was a significant difference between average scores at baseline and endline (p= 0.0012) and have added this to the text (L364-365).

Comment 13:

5. Presentation of results
It would be useful to present the resuscitation item results in a comparable manner to the structural indicators and the process of care indicators, perhaps ordered so that the care practices results are last, as process follows structure in the causal ordering.

Authors response: We have developed a table for the resuscitation score that is similar to the tables for the other domains as requested (new table 3)

Comment 14:
Each table could include the summary score for that domain. This would enable Table 3 to be deleted to draw further attention to Figure 2.

Authors response: We have removed the current table 3 and included the summary scores for each domain at the bottom of each table. We have included the overall summary score in the text (L259-265)

Comment 15:
I would suggest simplifying Figure 2 into a dot plot so that each hospital has three points on a line for the baseline, midline, and endline scores. This enables the same information to be presented in a more clear and compact form. Please label the X axis.

Authors response: We have revised figure 2 to include only baseline and endline scores (leaving out the midpoint scores) and have re-arranged the hospitals in order of the baseline score. We think this clarifies the figure considerably.

Debra Jackson (Reviewer 2): This is a clear and well written manuscript presenting findings from an evaluation of the KINC programme in South Africa. This project used a health systems approach rather than just a single focus approach which is much needed.

I have only a few minor suggested edits:

Comment 1: I feel it should be 'care 'FOR' small and sick newborns not 'OF' suggest correcting throughout manuscript, lines 72, 73, 76
Authors response: This has been corrected throughout the manuscript

Comment 2: line 88 should be 'in' the SA National PPIP, not 'on'
Authors response: the change has been made as suggested by the reviewer

Comment 3: line 95 insert 'and' between plan and inadequate
Authors response: the change has been made as suggested by the reviewer

Comment 4: line 127 clarify briefly the difference between intensive care and high care, e.g. intensive care has artificial ventilation and other advanced care?
Authors response: A brief note has been added to clarify the definition of intensive care (L 130-133)

Comment 5: line 140 please spell out MCWH as first use?
Authors response: the change has been made as suggested by the reviewer

Comment 6: the figure titled "Total Quality of Care Score" Is too busy and does not add anything as it is well covered in the text, I would delete this figure
Authors response: We have revised the figure in line with the comment from reviewer one, so that it is clearer and less busy.