Reviewer’s report

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.

Version: 1 Date: 17 Aug 2018

Reviewer: Hannah Leslie

Reviewer's report:

Dear authors,

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.

1. Study design
While the real-world nature of this intervention clearly precludes some randomized study designs, the pre-/post- designs provide 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. 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 1) 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. 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. 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. 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. 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.

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.

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. 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. 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. 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.

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.

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. 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. 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.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

No

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

No

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal.