Author’s response to reviews

Title: Exploring the effect of implementation and context on a stepped wedge randomized-controlled trial of a vital sign triage device in routine maternity care in low-resource settings

Authors:

Nicola Vousden (nicola.vousden@kcl.ac.uk)
Elodie Lawley (elawley91@hotmail.com)
Paul Seed (paul.seed@kcl.ac.uk)
Muchabaiwa Gidiri (fmgidiri@doctors.org.uk)
Umesh Charantimath (drumesh.charantimath@gmail.com)
Grace Makonyola (gmakonyola@gmail.com)
Adrian Brown (Adrian.brown@maternityworldwide.org)
Lomi Yadeta (kumsalomiyadeta@gmail.com)
Rebecca Best (rebeccambest@hotmail.co.uk)
Sebastian Chinkoyo (seb.chinkoyo@yahoo.com)
Bellington Vwalika (vwalikab@gmail.com)
Annettee Nakimuli (annettee.nakimuli@gmail.com)
James Ditai (j.ditai@liverpool.ac.uk)
Grace Greene (grace@secondmilehaiti.org)
Lucy Chappell (lucy.chappell@kcl.ac.uk)
Jane Sandall (jane.sandall@kcl.ac.uk)
Andrew Shennan (andrew.shennan@kcl.ac.uk)

Version: 1 Date: 18 Feb 2019

Author’s response to reviews:

Dear Editor,
Thank you for your consideration and positive reviewers’ comments on this manuscript. We have addressed each below and feel they strengthen the manuscript. In order to fully address the reviewer’s comments on mechanism of action and sustainability, we have added in additional analysis of nine focus group discussions held at six-nine months after implementation. These had previously been intended for a further manuscript, but we agree with the reviewers that they clarify the message of this study. In order to add the recommended sections, the demographic details of qualitative participants and direct quotes have been added as tables. We hope you view these addition favourably.

Kind Regards,

Andrew Shennan

Reviewer #1: Overall: This article was very interesting and well-written. Global health research will benefit from RCTs incorporating process evaluations to improve implementation and replication of effective interventions as highlighted in this paper. I am wondering why you decided to separate the qualitative results from the quantitative results, which nicely follow the RE-AIM framework, rather than weaving qualitative results within each framework piece, which would be more aligned with Table 1.

The qualitative results have now been integrated within the framework results as recommended:

“Implementation Fidelity

The average duration of implementation training across all facilities was 10.8 days. In total, 2747 HCP were trained, 61.1% of all those working in maternity services in those sites (range 16.5 in Kampala, Uganda to 89.2% in Zomba, Malawi, Table 1). Nine of the ten sites delivered all the key content of training. Freetown, Sierra Leone was the first to implement and less emphasis was placed on training senior staff, the background of device development and validation studies. Following challenges from senior staff in accepting device accuracy, this was emphasised in subsequent site training. Educational materials were translated (India, Ethiopia, Malawi, Haiti) and delivery was adapted to take into account locally available medications and referral structures. In India, all training was delivered by the research team rather than via CRADLE champions (87.1% trained). In Haiti, community HCP without formal training had a longer duration of training (approximately 2 days), using the same materials, to check understanding and confidence. The duration of training was longer in sites with a wider geographical spread or more challenging terrain (Mbale, Uganda; 18 days and Zomba, Malawi; 16 days) except in India, which was able to mobilise a larger local research team (10 days). External events influenced implementation in two sites, one of three tertiary hospitals in Cap Haitien, Haiti was closed at the time of implementation due to strike action of all staff, therefore key managers were trained, and remaining staff received training within two weeks of opening. In Ndola, Zambia implementation coincided with roll-out of alternative (un-related) training for some maternity
staff by the Ministry of Health. Implementation went ahead as planned for remaining staff and those that were unable to attend were trained by champions or the research team in the subsequent week.

Demographic details of the qualitative participants are shown in Table 2. Qualitative findings demonstrated that the majority of participants from all sites felt the training was adequate. Champions felt confident using the materials to orientate their colleagues. Recipients of training from champions were confident to use the VSA and also to orientate others. Champions felt confident using the materials to orientate their colleagues. Recipients of training from champions were confident to use the VSA and also to orientate others. A small minority of participants from the three sites that trained the fewest HCP (Addis Ababa, Kampala and Freetown) highlighted that training from the champions had been brief, that staff who were not trained took longer to learn and faced initial challenges with use, or that ongoing training may not be sustainable with staff turnover (quotes to illustrate in Table 3).

Reach

Overall, 3868 devices were delivered across 286 facilities. Four clusters recorded the proportion of women with BP measurement. All demonstrated a significant increase in measurements made after the intervention (usual care mean 79.2% (n=6093/7693) vs. intervention 97.6% (n=7800/7992); OR 1.30, 95% CI 1.29-1.31); Table 1). Prior to the intervention, 95% of facilities had access to at least one working BP machine. After the intervention, 100% had access, with better availability per HCP in all clusters. Participants from both clinics and hospitals in every cluster except Haiti, reported an increase in availability of equipment. The availability of equipment, and its ease of use, meant that more vital signs measurements could be done and faster, as staff did not spend time looking or waiting for equipment (Table 3).

Many participants reported that students and other allied HCP or volunteers would regularly help to take vital signs measurements with the device. More junior staff also took more vital signs measurements, where they would previously have referred the patient to other HCP for routine monitoring. This was reported to be due to greater confidence in their capacity to measure BP and interpret results. It was frequently commented that this made it more likely that women would have their vital signs measured (Table 3). In Haiti and Ndola, community HCP reported confidence and pride in being equipped and skilled to monitor vital signs in their community. This also led to more vital signs measurement in the community and earlier detection of abnormalities (Table 3). A minority of HCP reported that demand still outweighed supply, even though this was improved.

Adoption

The majority of sites reported rapid use of the device on all pregnant women. The reasons for rapid adoption differed according to site context. Sites with poor availability or poor-quality existing equipment (e.g. Kampala, Freetown, Mbale and Zomba) reported rapid use, irrespective of the different proportion of staff that were trained. Sites with adequate availability of
equipment prior to implementation (Gokak and Addis Ababa) elected to use the VSA in preference to other equipment citing ease of use, better accuracy and easier interpretation due to the traffic light alert which reduced the workload. This was true across all cadres of HCP from community volunteers to medical officers in hospitals.

Due to the stepped-wedge design, eight clusters reviewed use at six months post-implementation and three at 12 months. The majority of clinical areas were using solely the CRADLE VSA device at six months (73.1%; range 33.3% in Addis Ababa, Ethiopia to 90.2% in Ndola, Zambia, Table 1). Only 4.8% of clinical areas had chosen to use previously existing vital signs devices in preference to the CRADLE device. This was still reflected at 12 months (73.5% using solely the CRADLE device). A minority of sites reported barriers to adoption, the most frequent was the sensitivity of the VSA to movement and positioning, in some cases leading to mistrust of the accuracy of results. This was reported more frequently in sites with low fidelity (Freetown, Kampala and Addis Ababa). However, qualitative findings in Freetown suggest that active support from the champions or the research team resolved this concern, and this correlated with improved adoption compared to Addis Ababa and Kampala and over time (Table 1).

By the trial end, 4.6% (n=180) of VSA were reported to be broken. The most commonly reported reasons were failure of the battery, leaking of the valve in the pump or tears in the cuff. Many sites noted it was more robust than pre-existing equipment (Table 3). Very few CRADLE VSA were reported missing by the trial end (0.6% (n=23). Sites described self-directed systems of handover or registration to minimise this risk (Table 3).”

Background

* Citation numbers should be placed inside a sentence before the period, or before the comma notation.

This has been completed throughout.

Methods

* In row 9, you write the number of interviews and FGDs in each site. Could you also include N (total)? (You could bring up this info from the Results section (first paragraph in qualitative analysis).

This has been added as follows: “In total we conducted 36 interviews and 19 focus group discussions with 130 participants across the ten sites.”

* In row 26, you state that the comparison of referral rates before and after implementation is non-randomized. It seems you comparing the sites' referral rates to themselves? It would help if this line was clarified.
This has been clarified as follows “In each site a four-week period immediately prior to and three months after implementation were compared, therefore this is a non-randomized comparison.”

* For the logistic regression models, were there any control covariates used? Is this why a meta-regression was used? It is not clear in the paper why a meta-regression approach was used. I have only seen meta-regressions used for systematic literature reviews. Please explain further why this analytical approach was appropriate here.

We controlled for different event rates in each centre and trends over time in each centre, this is stated in the second sentence of the statistical analysis section. We did not control for other covariates because this was exploratory analysis and it was not obvious which to correct for. We chose this method of meta-regression because a two-stage approach (estimating effects first by centre and then trying to account for differences between centres) was likely to give a clearer description to the reader. This is a new field little guidance on standard methodology.

Results

* Large numbers such as those in row 40 or 47 would benefit from commas

These have been added.

* Was qualitative data analyzed by level of fidelity/core component implemented correctly by site or reach to explain these results? It would have helped if the qualitative results also followed the framework layout with the quantitative results.

Further text has been added to better integrate these results:

“Sites with poor availability or poor-quality existing equipment (e.g. Kampala, Freetown, Mbale and Zomba) reported rapid use, irrespective of the different proportion of staff that were trained. Sites with adequate availability of equipment prior to implementation (Gokak and Addis Ababa) elected to use the VSA in preference to other equipment citing ease of use, better accuracy and easier interpretation due to the traffic light alert which reduced the workload. This was true across all cadres of HCP from community volunteers to medical officers in hospitals.”

“A minority of sites reported barriers to adoption, the most frequent was the sensitivity of the VSA to movement and positioning, in some cases leading to mistrust of the accuracy of results. This was reported more frequently in sites with low fidelity (Freetown, Kampala and Addis Ababa).”

* Some of the information in the qualitative section should go under methods, particularly in the first paragraph.

This has been moved as recommended.

* Figures 1 and 2 - Shouldn't these titles reflect "change in odds"?
This should state the odds ratio. Both figure titles have been clarified as follows:

“Figure 1: Forest plot showing odds ratio for referral in individual clusters* in the intervention period compared to the control period.”

“Figure 2: Forest plot showing odd ratio for the primary outcome in individual clusters in the intervention period compared to the control period.”

Reviewer #2: I have read the paper with interest and I would like to suggest some work for improvement.

I feel it would be important to clarify on methodology how many domains are being analysed: 3 (Implementation, Reach and Adoption, as I can read on "Data collection", lines 47-48, page 9) or 6 (as you mention on "Outcomes", line 14-15, page 8) as both differ from what is indicated in Table 1 and actually is presented in results (Context - Fidelity - Reach - Adoption - Action (potential)), creating some confusion in the reader.

This clarity of these sections has been improved in line with the domains in Figure 2 (adapted from Table 1) and the results section:

“Clusters were ranked from highest to lowest on selected quantitative outcomes on implementation fidelity, reach and adoption (marked * in Figure 2). These were selected as the direction of benefit was clear, whereas the anticipated direction of change for outcomes on context and action were less clear (e.g. poorer availability of resources at the trial start may be associated with greater benefit from the intervention due to greater need, or less benefit due to inability to respond to abnormal vital signs). Outcomes under the same domain were averaged and converted to a possible range (0 to 1) to give each cluster a score for each domain analysed (implementation fidelity, reach and adoption).”

Despite the Table 1, the complexity of this work asks for a conceptual framework to help visualize and clarify the process of analysis, undertaken measures and outcomes, and expected results.

The manuscript has been improved by adding in a description of the hypotheses and the logic model to the introduction. Table 1 has been adapted into Figure 2 which better shows how the domains are informed by the RE-AIM framework and underpinned by evaluation of context and mechanism of action. This figure is shown below:

The text need some more detail on the following issues:

_the randomization process (e.g. were "stepping" randomized?) and on motivations as well.

The following has been added to improve this:
“The stepped-wedge design meant that clusters crossed over from control to the CRADLE intervention in one of nine steps at 2-monthly intervals over the 20-month trial duration. The order of steps was randomly allocated using a computer-generated sequence [21]. This design was chosen to minimise the risk of bias and show causality, should a significant effect of the intervention be demonstrated.”

_impact of time on effectiveness

The effect of the intervention on the primary outcome was adjusted for underlying time trends in each cluster. The text that describes this interaction in the statistical analysis plan has been improved to clarify this as follows:

“For the primary outcome in individual sites, the main analysis used logistic regression with generalised estimating equations and a population-averaged model. Adjustments were made for fixed centre effects (categorical) and separate fixed linear trends (continuous) in each centre to account for changes in the primary outcome over time [39].”

_the process dealing with the health workforce (HCP).

Regarding the last item, the CRADLE champions were random-selected individuals or teams? Were they asked to collaborate and participate in the intervention design (implementation design science). How was knowledge transferred from champions' training sessions to the other HCP?

Further information has been added to this section as follows:

“The device was delivered through a one-off interactive training session of CRADLE Champions. These were purposely selected health care providers (HCP) from each ward or facility in the trial cluster. They were selected prior to implementation, either as managers and/or as influential in their clinical area by the local research team. Interactive training sessions covered the use and maintenance of the device and suggested clinical management in response to abnormal vital signs using a combination of presentations, demonstration, practice and clinical scenarios. The CRADLE Champions were provided with posters, training manuals and a short, animated training film (sent by Bluetooth to smartphones). Two versions of the training materials were available, one for formally trained HCP working in facilities and one for HCP without formal training or working in the community. The CRADLE Champions then used these materials to provide ongoing training and support for use of the device in their clinical area. These components of the intervention and implementation were developed during a six-month feasibility phase with input from stakeholders [22].”

3 final questions for discussion:

_You mention "only 4.8% of clinical areas were not using the CRADLE device" (page 11, last line). Were they using another one or none at all?
This has been clarified as follows:

“Only 4.8% of clinical areas had chosen to use previously existing devices in preference to the CRADLE device.”

_In what way did the introduction of the device changed the local practices of registering and monitoring vital signs (e.g. BP)?_

The following section of the results under the subtitle ‘Reach’ has been improved with the addition of quotes to support the text in Table 3.

“All demonstrated a significant increase in measurements made after the intervention (usual care mean 79.2% (n=6093/7693) vs. intervention 97.6% (n=7800/7992); OR 1.30, 95% CI 1.29-1.31); Table 1). Prior to the intervention, 95% of facilities had access to at least one working BP machine. After the intervention, 100% had access, with better availability per HCP in all clusters. Participants from both clinics and hospitals in every cluster except Haiti, reported an increase in availability of equipment. The availability of equipment, and its ease of use, meant that more vital signs measurements could be done and faster, as staff did not spend time looking or waiting for equipment (Table 3).

Many participants reported that students and other allied HCP or volunteers would regularly help to take vital signs measurements with the device. More junior staff also took more vital signs measurements, where they would previously have referred the patient to other HCP for routine monitoring. This was reported to be due to greater confidence in their capacity to measure BP and interpret results. It was frequently commented that this made it more likely that women would have their vital signs measured (Table 3). In Haiti and Ndola, community HCP reported confidence and pride in being equipped and skilled to monitor vital signs in their community. This also led to more vital signs measurement in the community and earlier detection of abnormalities (Table 3).

_What are your findings considering long-term follow-up after stepping taking place?_

The following has been added to the discussion to draw together the findings on the sustainability of the intervention:

“Despite this, the validity of the CRADLE VSA as an accurate, robust, useful tool is maintained. Mixed-method follow up of use of the device at six months to one year after implementation is a strength of this study and supports the sustainability of the intervention. In addition, the proportion of devices that were broken or missing was lower than our sites report for previous existing equipment. Adoption was greater in sites that had higher proportions of HCP trained or more active CRADLE champions or local research teams to support the device. This suggests that these would be important factors for future scale-up.”

Reviewer #3: * This is an interesting paper of a vital signs triage devise used in routine maternity care in low resource settings. The authors conducted process evaluation to explore the
effect of implementation and context on a stepped wedge randomised controlled trial. The authors present a novel approach of evaluating implementation alongside effectiveness. Despite being a research study, and despite showing no correlation between the primary outcomes and different process evaluation domains, the findings present contextual and implementation factors important to similar maternal health interventions in low resource settings. Below, I have made a few observations aimed at strengthening the paper and mainly focuses on clarifying the study questions, logic model, mechanisms of action and mediators of action.

* There is a clear statement of purpose for the evaluation in the Introduction - "to describe the implementation of the intervention and the local contexts in which it was delivered and to determine whether differences in the effect of the intervention on the primary outcome can be explained." However, the study questions are not clearly presented. There is an attempt to mention them in the first column of table 1 but are not well articulated. Please list your study questions explicitly in the text (Introduction or Methods) and they should tie directly to the statement of purpose and Results. Justify why you chose to focus on these questions.

The objectives have now been clearly stated at the end of the introduction and directly link to Figure 2 (previously Table 1) and the results subheadings:

“These were chosen with the aim of exploring if and how this pragmatic intervention impacted on routine maternity care in a wide variety of settings:

- To evaluate whether the intervention was implemented as outlined in the protocol by describing the quantity and quality of training in each setting.

- To determine the reach of the intervention by evaluating the extent to which health care professionals and women were exposed to the intervention.

- To explore how the intervention was adopted into routine maternity care, whether this changed over time and the potential sustainability of this.

- To explore differences in context, implementation, reach and adoption between sites and determine whether they can explain differences in the effect of the primary outcome in different settings.

- To explore if and how the intervention impacted on routine maternity care across the facilities in each setting and identify possible reasons for this.”

* The MRC process evaluation guide recommends articulation of a programme theory (causal assumptions of complex interventions). There is no clear program theory constructed, tested, or refined. The authors have provided the logic model in the supplementary material. I recommend the authors to clearly articulate the programme theory in the actual paper (not as supplementary material - reflecting assumptions regarding the causes of the problem and how actions will produce change. Apart from the diagram, the authors should briefly explain (in a paragraph) the initial programme theory
for the CRADLE-3 VSA intervention and explain how it was elicited i.e. existing literature or the feasibility study.

We welcome the opportunity to expand on our theory and have added the following in the methods. The figure of the logic model has previously been published in the feasibility study described, we have added it to the manuscript as requested but request the editors to advise whether this is acceptable:

“It was hypothesised that better availability of equipment would improve the efficiency and capacity of HCP to monitor vital signs. It was also hypothesised that training would improve HCP understanding of when and how to measure vital signs and how to identify and manage pregnancy complications. The ease of use of the CRADLE VSA and the traffic light early warning system would mean that all cadres of HCP would be alerted to and have confidence to act on abnormal vital signs. Together, this would result in more women receiving more vital signs measurements, so abnormal results would be identified earlier and managed faster thus reducing maternal morbidity and mortality. These hypotheses were developed through field studies, stakeholder engagement and literature demonstrating need for improved access to equipment [31-34] training[35-37] and task-sharing[38, 39]. In addition, qualitative evaluation [28] and a mixed-methods feasibility study [22] (in three sites similar, but geographically distant, to our trial sites) determined that the device is robust and easy to use by any cadre of HCP, even those without extensive training, and that the training package and implementation strategy were acceptable and had potential to impact on clinical management (escalation and referral). During this feasibility study, several components of the intervention and implementation were refined and a logic model created (Figure 1) to identify the assumptions, processes and anticipated outcomes and the key areas for evaluation in this study.”

* Background: The health systems need to be described briefly or presented as a figure so that readers who are not familiar with the health systems in the Ten clusters, understand the provision of maternal health services (especially before the intervention) i.e. equipment used and the workforce employed.

The following general points have been added to the background to improve understanding of the health systems in this trial. Resource availability of the health systems is provided in Table 3:

“Yet, in LMIC, challenges such as inadequate numbers of trained health care providers (HCP) [27] and insufficient access to reliable, accurate, equipment to monitor vital signs [28-31] lead to delays in identifying women with pregnancy complications which contributes to preventable mortality and morbidity [4, 5].”

“This is important in LMIC where routine clinical tasks, such as vital signs measurement, are often undertaken by those with minimal training and community health workers also play a vital role in maternity care, often being the first point of contact and an essential link to clinical services[6, 32].”

Later is the Results, it will be good to show how differences in the health systems across the clusters affected implementation of the intervention
The following has been added as suggested:

“In Haiti, community HCP without formal training had a longer duration of training (approximately 2 days), using the same materials but spending greater time checking understanding and confidence.

Clusters that trained fewer staff tended to have multiple, very large facilities with high numbers of deliveries (Lusaka and Kampala). This is with the exception of Freetown, which was a smaller unit but trained fewer staff. This cluster was the first to implement, possibly demonstrating the learning curve of the research team. Qualitative findings demonstrated that the majority of participants from all sites felt the training was adequate (Demographic details of the qualitative participants are shown in Table 2). Champions felt confident using the materials to orientate their colleagues. Recipients of training from champions were confident to use the VSA and also to orientate others. A small minority of participants from the three sites that trained the fewest HCP (Addis Ababa, Kampala and Freetown) highlighted that training from the champions had been brief, that staff who were not trained took longer to learn and faced initial challenges with use, or that ongoing training may not be sustainable with staff turnover (quotes to illustrate in Table 3).”

* Page 6, line 42 - for the benefit of diverse readers please describe what is meant by vital signs. Vital signs can be described differently in different setups.

This has been added as below:

“The CRADLE Vital Signs Alert (VSA) accurately measures blood pressure (BP) and heart rate and calculates shock index (heart rate divided by systolic BP) [21-24]”

* Data collection methods - A variety of data sources are used. The nature of how 'contextual data' is collected during baseline and throughout the intervention is not clear (e.g. no interview guides, or observation tool). In table 1, methods and tools for data collection are clear for the other domains and not 'context' - it only says "measurement of resources at baseline and throughout the trial". This could be described briefly in the text.

Figure 2 (previously Table 1) has been improved to clarify the methods of measurement (quantitative evaluation of key resources and qualitative evaluation of the physical and cultural environment around measuring vital signs, clinical management, escalation and referral systems). The outcomes measured under context are described in the first paragraph under data collection and the limitations of these measures are explored in the last paragraph of the discussion.

“At the trial start, baseline data were collected from each facility on the distance from the nearest tertiary referral hospital, number of HCP working in maternity (doctors, nurses, midwives, clinical officers and community HCP in Ndola and Cap Haitien), availability of existing blood pressure equipment, blood transfusion services, intensive care beds and magnesium sulfate. These were selected as markers of health system context that were important and feasible to measure. This was updated a minimum of three times during the trial period. Major changes to
the political or physical environment such as infrastructure, staff retention and travel or extreme weather conditions were evaluated monthly.”

“In each site, we undertook semi-structured interviews (n=3-5) and focus group discussions (n=1) with HCP, three months after implementation. These explored the uptake of the intervention, its influence on clinical management and any unexpected consequences. In sites that implemented in the first 14 months of the trial, a further focus group discussion was undertaken at 6-9 months after implementation to explore whether influence on clinical management, escalation and referral systems changed over time and the sustainability of the intervention.”

There is no indication that additional data was collected to answer particular arising questions (during the iterative process of process evaluation).

Themes that arose during qualitative interviews and focus groups were added to the framework to be explored in more depth in subsequent sites. Whilst the diverse settings of this trial are a strength, the number of sites and resource constraints, as well as the simultaneous delivery alongside a stepped wedge trial design (with strict intervals for implementation) meant that collecting additional data was not feasible. This has been added as a limitation to the discussion.

“Whilst the diverse settings of this trial are a strength, the number of sites, resource constraints, and the simultaneous delivery alongside a stepped wedge trial design (with strict intervals for implementation) meant that there was limited capacity to collect additional data in response to early findings.”

* The evaluation identifies various contextual influences (resource availability, staff levels). These seem to be appropriate. However, other important contextual influences like infrastructure, weather and politics, are only mentioned in the data collection section (lines 40 - 47) but are not explained in the results how they affected implementation in the individual clusters.

The following has been added to the implementation fidelity section:

“External events influenced implementation in two sites, one of three tertiary hospitals in Cap Haitien, Haiti was closed at the time of implementation due to strike action of all staff, therefore key managers were trained, and remaining staff received training within two weeks of opening. In Ndola, Zambia implementation coincided with roll-out of alternative (un-related) training for some maternity staff by the Ministry of Health. Implementation went ahead as planned for remaining staff and those that were unable to attend were trained by champions or the research team in the subsequent week.”

And the context section:
“There were a number of external influences during the trial period, for example strike action in Kampala, Uganda and an earthquake outside of the research area in Haiti. However, sites reported minimal impact of these events on care provisions.”

* On several occasions, the manuscript has made reference to "Mechanisms of action" i.e. "The intervention and its implementation strategies were refined, and potential mechanisms of action explored (page 6 lines 19 - 24). The MRC process evaluation guide defines Mechanisms of action as 'how intervention activities, and participants' interactions with them, trigger change'. Mechanisms should be the causal pathway generating a particular outcome, and your results should focus on identifying these mechanisms and describing how context influenced them. Currently the results section is silent on Mechanisms. For instance, explore on how nurses got frustrated or motivated with the new VSA and how that affected outcomes, space restrictions and how it affected healthcare workers, relationships and it affected teamwork, interaction with women and the intervention.

Thank you. The following section on mechanisms of action has been strengthened and expanded considerably as follows:

“In addition to the mechanisms previously mentioned (better availability of equipment, ease of use and confidence of all cadres of HCP to measure vital signs), the increase in equipment and training meant it was no longer acceptable to not measure vital signs on every woman. Staff reported increased motivation and interest in vital signs measurements. Only one site (Mbale) reported this in a negative light, since measurement of BP on all women increased workload. The other sites reported a reduced workload as time taken to find equipment, measure vital signs and interpret results was reduced and this task could be undertaken by a wider number of HCP.

It was frequently reported that the intervention prompted HCP to do more investigations, more quickly. This was reported to be because the traffic light display alerted users to results outside the normal range, and HCP had more confidence in the results so were better able to make decisions. This finding was not dependent on the number or skill level of staff. A minority of participants opposed this view, stating that the management was unchanged, as vital signs were always measured and acted upon. This was most commonly reported by senior HCP working in better resourced environments. Even in this setting, benefit was still reported from the traffic light alert in aiding communication between HCP.

The majority of sites also reported that the alerts were easily understood by women and untrained staff such as ambulance drivers. This was beneficial in conveying the need for management or referral, especially in sites where this was reported to be a key barrier to care (Gokak, Ndola, Zomba, Harare). Some sites reported that increased awareness of vital signs in the community resulted in increasing demand for measurements to be done (Table 3).

The impact on referrals differed between sites. Overall, 3.7% (n=2784/74,828) of women seen in peripheral maternity facilities were referred to higher level care in the control period compared to 4.4% (n=3212/73,371) in the intervention period (OR 0.89; 0.39-2.05) (data for nine sites that were able to collect denominator). However, the majority of sites demonstrated a small but
significant reduction in referrals with a single site (Gokak) demonstrating a 16-fold increase (Figure 3, Table S1). Qualitative findings suggest the increase in Gokak was a result of increased community monitoring, increased confidence in peripheral HCP to detect abnormal vital signs and convince women to attend, alongside rigorous adherence to referral protocols from rural health posts (subcentre) to primary care centres, meaning all women with asymptomatic anaemia triggering a yellow light were referred. This is in combination with an effective ambulance system, and further systems in place to cope when ambulance services were delayed, to transfer patients from primary care clinic to hospital when acute complications were detected. Therefore, the wide geographical distance (mean 74km from peripheral clinic to tertiary hospital) of this site did not impede delivery of care.

In contrast, Haiti reported no change in the number of referrals but that abnormal vital signs were detected and referred faster by using the traffic light alerts to convince women to attend, where cultural acceptability and perceived quality of hospital care was a barrier. However, despite the relatively small cluster size (mean 14km from peripheral clinic to tertiary hospital), the qualitative data indicated that the lack of ambulance service or funds (personal or within the health care facility) to pay for transport led to long delays contributing to morbidity and mortality, irrespective of the capacity to monitor and escalate care peripherally.

Differing acceptability of referrals and the relationship between peripheral and tertiary facilities arose as important contextual themes that may have facilitated or impeded action from the intervention. For example, HCP in Lusaka (significant benefit of the intervention) described an existing mechanism for constructive feedback on referrals between facilities, which was aided by the introduction of uniform monitoring equipment. In comparison, HCP from both peripheral and tertiary facilities in Zomba (no benefit of intervention) described negative concerns about referral, such as a lack of system to alert the recipient hospital of the pending transfer resulting in patients being refused admission, which potentially deters future referrals. HCP in peripheral facilities in Mbale (no benefit of intervention) reported that referrals were reduced following the intervention, since pre-eclampsia could now be managed in the community, which was encouraged by the tertiary facility.”

* Qualitative data - there is a clear description of what qualitative data was collected and how they collected it. This has been well laid down in the Results section (qualitative analysis). However, this section runs short of useful direct quotes as is expected of qualitative research to substantiate our statements.

Table 4 has been added with direct quotes to support all the qualitative analysis in the results section (due to limited word count).

* Mediators of action - Table 1 has defined each of the domains in the results i.e. context as the setting for implementation for each cluster. However, there is no definition of what the evaluation refers to as 'mediators of action'. I recommend to the authors to define this. Other literature looks at this as 'intermediate processes which explain subsequent changes in outcomes'. It is therefore unusual for this evaluation to refer 'mediators of action' of such a complex intervention just as referrals of maternity visits. Please furthers explore what were mediators and how they influenced outcomes.
The definition in the table has been changed to mechanism of action in keeping with the text of the manuscript. The clarity of the methods section has been improved to describe these measures:

“In order to evaluate the ways in which the intervention, and participants interaction with it, may trigger change (mechanisms of action) that were identified in the feasibility study[38], the number of women attending maternity services, the proportion that had their blood pressure measured and the proportion referred to higher level care were measured for a four-week period immediately prior to implementation and three months after implementation.”

The addition to the results section has been described in the previous comment.

* Page 10, line 43 - deliveries at home. Please explain what is meant by ‘systematically collected’. How data on deliveries at home were collected in the seven clusters? You might need to explain this in the methods.

The following has been added to the methods to clarify this point:

“The number of deliveries in each cluster was collected by review of facility registers and routine reporting. Community deliveries were captured in seven of the ten sites through a variety of methods such as household visits from community health workers in India and monthly reporting meetings with traditional birth attendants in Haiti (three sites did not routinely record deliveries that occur outside of facilities).”

* In view of the CRADLE VSA that went missing (23), those that got broken down (180), and clinics that were not using them six months post intervention; what is your comment on the sustainability of the equipment? Based on the reasons for missing or breaking down of the machines, and based on your experience in implementing the trial, will you be able to provide recommendations on how to prevent this? i.e. issues of maintenance and security?

The following has been added regarding the sustainability of the device to the adoption results,

“By the trial end, 4·6% (n=180) of VSA were reported to be broken. The most commonly reported reasons were failure of the battery, leaking of the valve in the pump or tears in the cuff. Many sites noted it was more robust than pre-existing equipment (Table 3). Very few CRADLE VSA were reported missing by the trial end (0·6% (n=23). Sites described self-directed systems of handover or registration to minimise this risk (Table 3).”

The following has been added to the discussion:

“Despite this, the validity of the CRADLE VSA as an accurate, robust, useful tool is maintained. Mixed-method follow up of use of the device at six months to one year after implementation is a strength of this study and supports the sustainability of the intervention. In addition, the
proportion of devices that were broken or missing was lower than our sites report for previous existing equipment. Adoption was greater in sites that had higher proportions of HCP trained or more active CRADLE champions or local research teams to support the device. This suggests that these would be important factors for future scale-up.”

* The evaluation shown no correlation between process measures within domains and no correlation between individual domains and the primary outcome. However, there are other process outcomes that improved as a result of the intervention and these outcomes varied across the clusters. Such variations provides an opportunity to describe contextual and other influences on the process outcomes. I would like to see this focussed more on discussing the connections between context and mechanisms. You have the opportunity to not just to focus on correlation between the domains and primary outcomes - but to delve into which are most important contextual influences to influence (or not influence) process outcomes. This is the core of process evaluations - explains what works for whom and in what conditions.

Relate the two e.g. number trained to number of staff, uptake of the device by cadre/number of staff/size of hospital

The relationship between process measures and context has been strengthened throughout. For example:

“This was reported to be because the traffic light display alerted users to results outside the normal range, and HCP had more confidence in the results so were better able to make decisions. This finding was not dependent on the number or skill level of staff. A minority of participants opposed this view, stating that the management was unchanged, as vital signs were always measured and acted upon. This was most commonly reported by senior HCP working in resourced environments. Even in this setting, benefit was still reported from the traffic light alert in aiding communication between HCP.”

“This is in combination with an effective ambulance system, and further systems in place to cope when ambulance services were delayed, to transfer patients from primary care clinic to hospital when acute complications were detected. Therefore, the wide geographical distance (mean 74km from peripheral clinic to tertiary hospital) of this site did not impede delivery of care.

In contrast, Haiti reported no change in the number of referrals but that abnormal vital signs were detected and referred faster by using the traffic light alerts to convince women to attend, where cultural acceptability and perceived quality of hospital care was a barrier. However, despite the relatively small cluster size (mean 14km from peripheral clinic to tertiary hospital), the qualitative data indicated that the lack of ambulance service or funds (personal or within the health care facility) to pay for transport led to long delays contributing to morbidity and mortality, irrespective of the capacity to monitor and escalate care peripherally.”

“Differing acceptability of referrals and the relationship between peripheral and tertiary facilities arose as important contextual themes that may have facilitated or impeded action from the
intervention. For example, HCP in Lusaka (significant benefit of the intervention) described an existing mechanism for constructive feedback on referrals between facilities, which was aided by the introduction of uniform monitoring equipment. In comparison, HCP from both peripheral and tertiary facilities in Zomba (no benefit of intervention) described negative concerns about referral, such as a lack of system to alert the recipient hospital of the pending transfer resulting in patients being refused admission, which potentially deters future referrals. HCP in peripheral facilities in Mbale (no benefit of intervention) reported that referrals were reduced following the intervention, since pre-eclampsia could now be managed in the community, which was encouraged by the tertiary facility.”