Breath REcognition Aid to Health Experts (BREATHE) study 2019

Results report

The funder of the study is the Philips Foundation. The Philips Foundation is a registered charity, established in 2014 as the central platform for Philips' Corporate Social Responsibility activities.

Prepared by Charlotte Ward, Ann-Sophie Stratil, Kevin Baker, April 2020

Front cover: village in SNNPR, Ethiopia

Photo credit: © Malaria Consortium

Contents

List of abbreviations

ARIDA	Acute respiratory infection diagnostic aid
ARI	Acute respiratory infection
BREATHE	Breath recognition aid to health experts
BPM	Breaths per minute
CHW	Community health worker
FGD	Focus group discussion
HEW	Health extension worker
ICC	Intra-class coefficient
iCCM	Integrated Community Case Management
IMCI	Integrated Management of Childhood Illness
IMNCI	Integrated Management of Newborn and Childhood Illness
LoA	Limits of agreement
MoH	Ministry of Health
PDP	Pneumonia Diagnostics Project
RR	Respiratory rate
SNNPR	Southern Nations, Nationalities, and Peoples' Region (Ethiopia)
U5	Under 5 (years old)
WHO	World Health Organisation

Summary

Background

Pneumonia is the leading infectious cause of death in under-five children. As per current World Health Organisation (WHO) guidelines, community health workers across low-resource settings currently count respiratory rate as a proxy sign for pneumonia. However, community health workers often find it difficult to accurately count breaths. New, automated respiratory rate counters offer a potential solution. To introduce new tools, their performance must first be validated against a robust reference. Currently, there is no gold standard.

The aim of this study was to build evidence around a new manual video annotation tool as a reference standard for assessing respiratory rate in children under 5 with cough or difficulty breathing.

Methods

This was an interrater reliability study to evaluate agreement between reviewers assessing the respiratory rate of children, using a manual video annotation tool. Video data had been collected in two previous studies conducted by Malaria Consortium [1, 2]. The study was set in Hawassa, Ethiopia. Data were collected between April and September 2019.

The new tool allowed reviewers to manually annotate certain breaths, uncertain breaths and distortions (non-breath movements and other interruptions to the normal breathing such as crying). The tool had functionalities including slowing down playback time, zooming in and out, adjusting brightness and moving back and forth along the video timeline. Based on the annotations, the respiratory rate was calculated as breaths per minute. Respiratory rate can be obtained in several ways, depending on the type of breath and time period considered in the calculation. For the primary results presented here, respiratory rate was calculated by considering certain breaths during calm periods.

A panel of ten reviewers were recruited. Reviewers were medical staff with at least two years’ paediatric experience and experience in manually counting respiratory rate, and received a four-day training on using the tool. Fifty one videos of children were selected for assessment and included children under 5 with cough or difficulty breathing ensuring equal representation of the three clinically relevant age groups 0 to 2 months, 2 to 12 month and 12 to 59 months. Each video was assessed by a random group of five reviewers of the larger reviewer panel.

Reliability was measured by intra-class correlation coefficient (ICC) for continuous respiratory rate as the primary outcome, Fleiss’ kappa for breathing status (classification normal/ fast based on WHO guideline criteria), proportion of videos where all five reviewers agreed on breathing status and range in respiratory rates assessed per video. Results were stratified by age group of children and average distortion period per video marked by reviewers.

Additionally, qualitative data from focus group discussion with panel members was evaluated using thematic analysis to assess usability and acceptability of the video annotation tool of the new tool as a reference standard.

Results summary

Overall, agreement on continuous respiratory rate between five reviewers was good (ICC = 0.93 [95% CI: 0.89; 0.95]). This corresponded to an average respiratory rate range of 9.19 [95% CI: 7.19; 11.19] bpm between the rate counted by the highest and lowest reviewer. There was substantial agreement on breathing status between the five reviewers (kappa = 0.71) which corresponded to all five reviewers agreeing on breathing status for 70% of the videos. Agreement as per Bland-Altman for the same two reviewers (n=15) was comparable to a previous study [2], with a bias of 0.2 bpm and limits of agreement -6.62 [95% CI: -7.6; -5.6] bpm to 7.02 bpm [95% CI: 6.0; 8.0] bpm.

Agreement measures differed between age groups and were generally lower in younger age groups: ICC was 0.87 [95% CI: 0.77; 0.94] for the age group 0 to 2 months and in comparison 0.94 [95% CI: 0.89; 0.98] for the age group 12 to 59 months. Agreement on breathing status was moderate in the age group 0 to 2 months (kappa = 0.48), while substantial in the age group 12 to 59 months (kappa = 0.80). Agreement between five reviewers was higher in videos with less distortion. Younger children showed a significantly higher degree of distortion than older children. These findings suggest that the interrater reliability of the tool depends on the distortion observed in the videos which is correlated with the age group of the child.

Qualitative assessments supported these findings: reviewers pointed out that videos with lots of distortions, movements and uncertain breaths were difficult to annotate with age being a factor influencing this difficulty. Identifying breaths in restless and crying children was challenging as the abdomen tends to become rigid and breathing could not be seen. However, annotators perceived the respiratory rate to be more accurate with the annotation software than using manual counting, as it allowed them to distinguish between normal breaths, uncertain breaths, distortion and movement. They also mentioned that tool functionalities, like slowing down, changing colour adjusting brightness helped them in distinguishing between these different kinds of breaths and movements.

Discussion

Our results are in line with findings from previous studies investigating interrater agreement on respiratory rate using manual methods and videos. Our findings indicate that reliability of respiratory rate derived from the new video annotation tool are influenced by child agitation and age: videos with lower levels of distortion and of older children show higher reliability than videos with high levels of distortion and of younger children. A robust reference standard should provide reliable measures under those circumstances as new diagnostic aids will need to be validated under real life settings. Acceptable limits of agreement are currently under review by the global community, and consensus on acceptable levels of accuracy and reliability need to be reached before final conclusions can be drawn.

Background and rationale

Pneumonia is the leading infectious cause of death in under-five children. As per current World Health Organisation (WHO) integrated management of childhood illness (iCCM) guidelines [3], community health workers (CHWs) across low-resource settings currently count respiratory rate (RR) as a proxy sign for pneumonia. However, there are known challenges to accurately counting RR because it is hard to manually define what is and is not a breath, it is easy to lose count, the child may be agitatedand there may be external distractions in the environment. Misdiagnosis of suspected pneumonia is common and can lead to over and under treatment with antibiotics and potential death.

New, automated RR counters offer a potential solution. To introduce new RR counters, their performance must first be validated. Developing a robust reference standard for evaluating the performance of new RR counters is challenging and there is currently no gold standard.

Previous studies have used contemporaneous counting by expert clinicians [4], retrospective review of video recordings by a panel of experts, [5-7] and other devices including capnography as RR reference standards [4].

The Breath Recognition Aid to Health Experts (BREATHE) study aims to address the global lack of a reference standard to accurately validate new RR devices for children under the age of 5. The study focus is to build the evidence base around the interrater reliability of manual video annotation as a reference standard for counting RR in children under 5.

Research team

- Senior Research Specialist - responsible for technical oversight, study conceptualisation, protocol writing, analysis oversight and report writing.
- Research Advisor - responsible for technical oversight, study conceptualisation, protocol and report review.
- Epidemiologist – responsible for quantitative and qualitative data analysis and report writing.
- Senior Programme Officer – responsible for study implementation in Ethiopia and quality assurance.
- Project manager – responsible for study implementation, team management, ethics and logistics in Ethiopia.

Methods

Study design

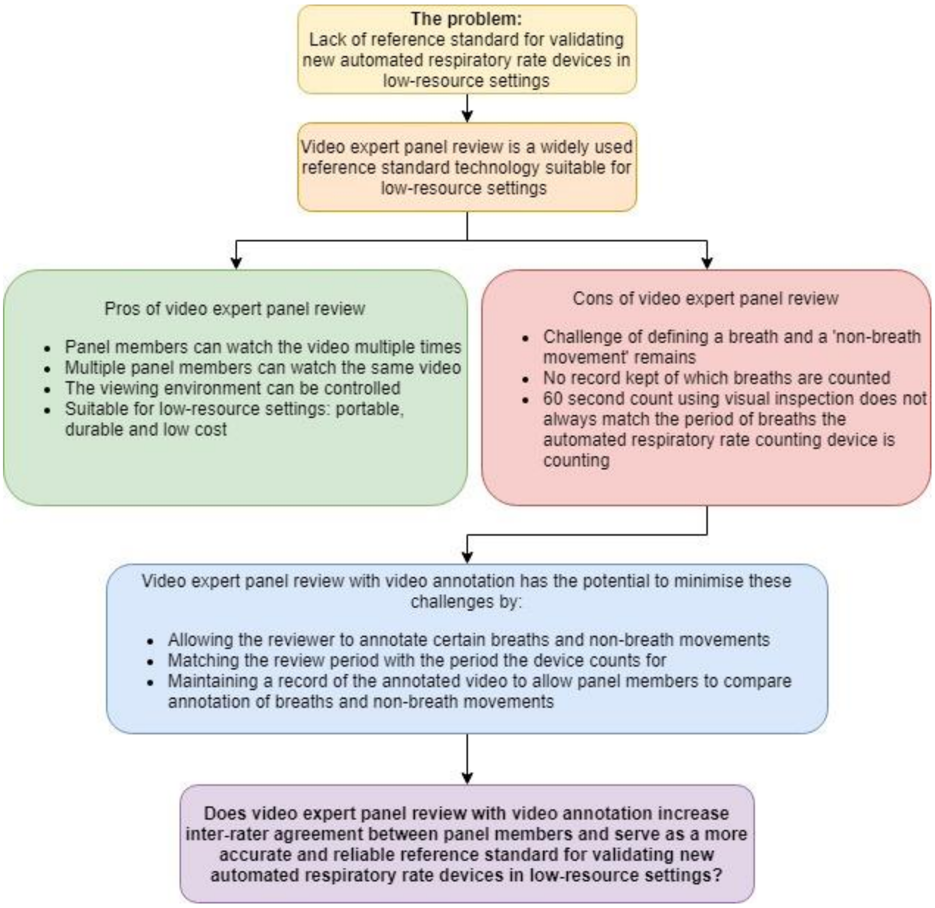
This was an interrater reliability study assessing the agreement between multiple reviewers on RR measures derived from a new video annotation tool. The assessment used videos of children under 5 with suspected pneumonia that had been collected in two previous studies conducted by Malaria Consortium[1]. The new video annotation tool facilitated the annotation of certain, uncertain and distorted breaths and had functionalities including slowing down playback time, zooming in and out, adjusting brightness and moving back and forth along the video timeline. Each video was annotated by five reviewers from a video expert panel of ten. RRs were derived from these annotations. Usability and feasibility of the new manual video annotation tool was assessed by focus group discussion (FGD) with panel members. The study design was agreed and finalised at the 5-day protocol design workshop that was held in Hawassa in February 2019. This meeting was attended

[1] ARIDA Diagnostic Agreement study funded by “la Caixa” Banking Foundation and Pneumonia Diagnostics Project funded by Bill and Melinda Gates Foundation.

by the Research Advisor, Senior Research Specialist, Senior Programme Officer and Project Manager and a representative from Philips Foundation and allowed for the draft protocol and data collection tools to be reviewed and finalised.

Conceptual framework

The framework below (Figure 1) outlines the rationale for building the evidence based for a video expert panel review with a video annotation tool.



The problem:
Lack of reference standard for validating new automated respiratory rate devices in low-resource settings

Video expert panel review is a widely used reference standard technology suitable for low-resource settings

Pros of video expert panel review

- Panel members can watch the video multiple times
- Multiple panel members can watch the same video
- The viewing environment can be controlled
- Suitable for low-resource settings: portable, durable and low cost

Cons of video expert panel review

- Challenge of defining a breath and a 'non-breath movement' remains
- No record kept of which breaths are counted
- 60 second count using visual inspection does not always match the period of breaths the automated respiratory rate counting device is counting

Video expert panel review with video annotation has the potential to minimise these challenges by:

- Allowing the reviewer to annotate certain breaths and non-breath movements
- Matching the review period with the period the device counts for
- Maintaining a record of the annotated video to allow panel members to compare annotation of breaths and non-breath movements

Does video expert panel review with video annotation increase inter-rater agreement between panel members and serve as a more accurate and reliable reference standard for validating new automated respiratory rate devices in low-resource settings?

Figure 1 Conceptual framework of developing a new reference standard for validating new automated respiratory rate devices

Study setting

Video expert panel members were stationed in an office in Hawassa, SNNPR, Ethiopia to review and annotate selected videos. Malaria Consortium has strong relationship with the SNNPR health bureau and were involved in a previous ARIDA project on automated pneumonia diagnostic devices. This study is significant to SNNPR and Ethiopia which have an extensive health extension programme with thousands of health extension workers manually counting RR.

Quantitative data were collected in April 2019, and qualitative data collection took place in April and September 2019.

Objective and outcomes

Study objective: To measure the interrater reliability of RR measures derived from a manual video annotation tool.

Primary outcome:

- The agreement between a group of five reviewers assessing the RR of selected subjects using a video annotation tool as measured by intra-class correlation coefficient (ICC)

Secondary outcomes:

- The agreement between a group of five reviewers on RR classification of selected subjects using a video annotation tool as measured by the Kappa statistic
- The mean RR range between the maximum and minimum RR measured by a group of five reviewers using a video annotation tool
- The proportion (%) of videos with agreement in RR classification between a group of five reviewers using a video annotation tool
- The agreement between two randomly selected reviewers using a video annotation tool as measured by mean difference in RR and limits of agreement (Bland-Altman analysis)
- The agreement between two of the same reviewers using a video annotation tool as measured by mean difference in RR and limits of agreement (Bland-Altman analysis)
- Average standard time taken per video
- Average standard distorted period per video
- The usability and acceptability of the video annotation tool to the video expert panel as analysed by focus group discussion

Video annotation tool

The video annotation tool was developed by a Senior Scientist at Philips to annotate videos of child's chest movements and other non-breath movements or distortions e.g. crying. Hereafter, the term 'distortion' includes non-breath movements and other interruptions to the normal breathing such as crying. The tool reports RR as breaths per minute (bpm) and allows the user to define start and end points, change the speed of playback and zoom levels, change brightness and mark breaths (normal and uncertain) and distortions. As per the WHO IMCI guidelines it is important that the child is calm for the duration of the RR count to prevent the counter from misinterpreting a non-breath movement as a breath.

If the measurement duration is not exactly 60 seconds, the tool automatically adjusts the RR so that it reflects the number of breaths in one minute. The tool also accounts for incomplete breath cycles at the start or end of the assessment in the calculation of bpm.

If the child was fully calm during the measurement and all breaths were certain, there will be a single RR measure. However, if any of the breaths within the measurement time is marked as not-calm or uncertain[2], the video annotation tool will calculate two numbers.

"Bpm upper": aims to calculates a measure reflecting the highest possible count that a clinician might have, by interpreting the WHO guideline 'count only when the child is calm' by excluding any distortion breaths and durations of time marked with distortion. Uncertain breaths and their associated duration are included to ensure the highest possible RR is reached.

"Bpm lower": aims to calculate a measure reflecting the lowest possible count that a clinician might have, by following the rule 'count for a full minute'. This excludes uncertain breaths and includes breaths during distortion. Uncertain breaths are excluded to ensure the lowest possible RR is reached. Distorted breaths are included because the clinician has to count for a full minute.

The full user manual can be found in appendix 1.

Study population

Video expert panel sample

Ten video expert panel members were recruited according to the following criteria:

1. **Medical experience**
   a. Minimum first degree, health officer or nurse
   b. 2 years working in health facility
   c. 2 years paediatric experience with experience counting RR or IMCI trained
2. **English proficiency**
   a. Excellent written and spoken English
   b. Certificate desired
3. **IT capacity**
   a. Certificate in basic computers

The recruitment process included an interview in English and a pre-test to assess the candidates' IT proficiency. For each video to be reviewed, a random set of five reviewers was selected from the larger video expert panel consisting of ten reviewers by simple random sampling using an online random number generator. A random selection of five reviewers participated in the FGD.

Video sample

A video sample was selected from a total of n=146 videos from previous studies: n=98 videos were from the previous ARIDA project (conducted at selected hospital in Ethiopia, i.e. Saint Paul's Hospital and Millennium Medical College in Addis Ababa, Ethiopia) and n=48 videos were from the previous PDP study (conducted at health facilities in Uganda, Ethiopia and South Sudan).

During the respective studies, the children for video documenting had been selected based on the following criteria (Table 1).

[2] Very shallow, incomplete cycle or difficult to judge

Table 1 Eligibility criteria from previous studies

Criteria	**ARIDA study**	**PDP study**
Inclusion criteria	1. Children aged 0-59 months with parent or guardian consent 2. For those aged 2-59 months, child must also have had cough or difficulty breathing	1. Children aged 0-60 days 2. Children aged 2-59 months with a cough and/or difficulty in breathing
Exclusion criteria	1. General danger signs [8] 2. Signs of severe pneumonia [9] 3. IMNCI pink referral signs for severe disease[3], 4. In-patient children who were managed by barrier nursing (such as severe burns, child with neutropenia, severe infectious diseases) and those not eligible for research procedures as advised by the supervising clinician, 5. Parent or guardian's age less than 16 years 6. No parent or guardian consent 7. Device manufacturer safety exclusion criteria [10]	1. Children with an illness of > 2-week duration 2. Children exhibiting one or more of the IMCI danger signs (severe dehydration, agitation, inconsolable, neck stiffness, active convulsions or fits, unconscious or lethargic, not breastfeeding, and vomiting everything) 3. Children with severe burns, with neutropenia, or with a severe infectious disease, 4. Children deemed ineligible as advised by the supervising clinician

ARIDA=ARIDA Diagnostic Agreement, IM(N)CI=Integrated Management of (Newborn and) Childhood Illness, PDP=Pneumonia Diagnostics Project

For the BREATHE study, n=51 videos were selected from this pool of n=146 videos via stratified random sampling by video source, i.e. ARIDA study and PDP study. To ensure equal representation of all age classes, selection was conducted via stratified random sampling by age bracket (i.e. 0-<2 months; 2-<12 months, 12-59 months). Only videos and annotation periods that fulfilled the following eligibility criteria were included (Table 2). For all videos, an annotation period was pre-defined to ensure that all reviewers watched the same video sections and those video sections were not too distorted.

Table 2: Eligibility criteria for present study

Criteria	BREATHE study
Inclusion criteria	1. The full chest and belly is visible for the duration of the video 2. The camera remains still for the duration of the video 3. There is at least 60 seconds of footage 4. The video is sufficiently bright 5. The video is free from external distractions 6. The child is not hiccupping during the video 7. The child is calm for the duration of the video
Exclusion criteria	1. The child cried for a significant proportion (30 seconds or more) of the video

Sample size

Sample size calculations for the video sample were based on the primary outcome ICC. Sample size calculation is following a method proposed by Bonett (see Equation 1). The formula estimates the required number of videos k based on the desired width ω of the confidence interval of assumed outcome estimate ρ. Z is the z score derived from the desired confidence interval, α is the desired type I error and n is the number of reviewers. Bonnet also suggests a correction of k + 1.

Equation 1 Sample size calculation formula

$$k = \frac{8z_{\alpha/2}^2(1-\bar{\rho})^2\{1+(n-1)\bar{\rho}\}^2}{w^2 n(n-1)}$$

Assumptions made are: type I error was set to a standard α=0.05; z score was set to 1.96 derived from a desired 95% confidence interval (95% CI); width of 95% CI was set to ω=0.2 allowing for a precise estimate of ρ; and number of reviewers was set to n = 5 for practicality reasons. Assumed ρ was estimated to be a conservative 0.7 based on the outcomes of a previous study [11].

Under these assumptions and accounting for a correction of k + 1, the required number of videos to be watched by a random set of n=5 reviewers was k =51 videos, allowing for an ICC estimate of ρ = 0.7 with precision of a 95% CI with a width of ω = 0.2 and a type I error of 5%.

Training and competency testing

Ten panel members were trained for four days by the Research Advisor and Senior Programme Officer who had previously been trained on the annotation tool by the Senior Scientist at Philips.

The training consisted of the following modules:

1. Introduction to Malaria Consortium and the BREATHE study
2. Introduction to the pneumonia context in Ethiopia and the existing RR counting aids
3. Refresher training on RR counting
4. RR counting exercise on a WHO training video with defined RR
5. Manual RR counting test
6. Challenges around RR counting reference standards
7. Introduction to the video annotation tool
8. Practice using the video annotation tool
9. Group discussion, comparing annotations with a reference video including five videos previously annotated by a Senior Scientist at Philips
10. Iterative practice on the video annotation tool
11. Annotation software test using two previous ARIDA pre-test videos (one with and one without distortion).

Assessments were conducted before data collection was started. All ten reviewers were within +/-3 bpm of each other during the manual RR counting test after two attempts.

For the annotation software test, all ten reviewers were within +/-2 bpm of the “gold standard”[4] annotation for a video without distortion. For the video with distortion, 8 out of 10 reviewers were within +/- 2 bpm of the “gold standard” annotation (RR lower). All 10 reviewers were within +/- 3 bpm for the bpm lower[5] and 9 out of 10 reviewers were within +/-3 bpm for the bpm upper[6]. See Appendix 2 for full results.

[4] Annotation of the Senior Scientist at Philips

[5] See section ‘Video annotation tool’ for definitions

[6] See section ‘Video annotation tool’ for definitions

Ethical approval

The SNNPR Regional State Health Bureau Health Research Ethics Committee gave ethical approval for the study on 4th April 2019 (ref (PN6/9/32080). Research assistants obtained written consent for observations and interviews from each reviewer. All videos used in the study were anonymised using a unique identifier code (UIC). No patient identifiable information (name, age) was directly linked to the video. Caregivers had given consent for the data to be used in future studies.

Data collection

Pre-test

Selected videos were pre-loaded into the video annotation tool on each reviewers' laptop by the Senior Programme Officer. The Senior Programme Officer created a schedule to assign which videos each reviewer would review each day to ensure they were not reviewing the same video at the same time on the same day, to prevent conferring.

Video review

The office where reviewers sat was light, airy and comfortable.

Video reviewers spent two to three hours per day reviewing and annotating videos. It was expected that each video would take around one hour to thoroughly review and annotate, using the following steps:

1. Open the video annotation tool and the pre-loaded video.
2. Ensure the time band at the bottom is zoomed to 1-second intervals.
3. Watch the full video at normal speed without annotating to decide:
    a. Where to focus on the child's chest/abdomen for the easiest view of breathing. This should be an area where there is regularity in breathing (where possible)
    b. Whether to toggle the brightness [c]
    c. The speed of playback
    d. The zoom of the video
4. Take particular notice of:
    a. Period(s) of non-breath movements
    b. Period(s) of breath movements
    c. Sounds (if applicable)
5. Complete the video set-up checklist to log the video configuration used for annotation.
6. Slowly review the video and mark all breaths (certain, uncertain and breaths during movement) at the point where the chest is fully expanded. Zoom into the time box and move the mouse along, clicking or pressing [N] to add a breath. Review the video as many times as necessary to correctly mark the certain breaths.
7. Re-review the video and re-annotate any uncertain breaths by selecting a breath with [2].
8. Re-review the video and mark any motion with [X].
9. Once the reviewer had reviewed his/her assigned videos for that day, s/he informed the project manager who checked that the assigned videos had been fully annotated and recorded this in the schedule.

Data extraction

The Senior Programme Officer extracted data from the video annotation tool at two time points per video:

- After all breaths have been marked [N] (phase 1) – copy the data from the 'golden' file (.anno) into excel spreadsheet
- After all certain [N], uncertain [U] and periods of distortion [X] have been marked (phase 2) - copy the data from the 'golden' file (.anno) into excel spreadsheet
- Save the excel spreadsheet, 'golden' file (.anno) and the .csv annotated file and the 'video_reduced' file (.mp4) for each annotated video, save them in a folder for each child UIC in their master folder.
- The project manager then copied in the reference time file 'meas_time' (.txt) in the format e.g. 1.000000, 61.000000 (where the start time is 1 second and the end time is 61 seconds). Each video had a unique time file that was copied over to the folder containing the .anno and the .mp4 file.

Time taken

Time taken was self-timed by the reviewers using a stopwatch.

Focus group discussions

Two FGDs with members of the video expert panel who annotated videos for the study were conducted to gather information on usability and acceptability of the video annotation tool. The topic guides were developed using a comprehensive conceptual framework of acceptability of healthcare interventions [12]. FGDs took place in Malaria Consortium Hawassa office and were audio recorded. They were facilitated and transcribed by a trained interviewer. An assistant researcher took notes during the first discussion. FGDs were conducted in Amharic and one research assistant was subsequently responsible for translating and transcribing the FGD verbatim. The first FGD was conducted on 17th of April 2019 during a week of data collection for the study. The five video expert panel members who had annotated videos that day were selected for participation (i.e. based on availability) in the FGD. The second FGD was conducted on 25th of September 2019 after analysing the first FGD and noticing some information gaps. This time, the five video expert panel members not included in the 1st FGD participated.

Topics of interest for discussion included:

- Perceived effectiveness of the training
- Perceived effectiveness of the tool
- Confidence in using the tool
- Ease of use of the tool:
    - Learning to use the tool
    - Opening the tool, navigating to the video
    - Marking breaths and motion
    - Changing magnification, speed, brightness
    - Interpretation of the results
- Time taken to complete the annotation
- Improvements to the tool and whether they would recommend the tool

Quality assurance, supervision and monitoring

The Senior Programme Officer spent the duration of data collection with the video expert panel members, collecting data and ensuring that video expert panel members were following the SOPs. She notably ensured that the video expert panel members were working independently.

The Senior Programme Officer sent data on a daily basis to the Epidemiologist and the Senior Research Specialist based at Malaria Consortium's HQ who were doing spot checks to ascertain the quality of the data.

Data analysis

Quantitative analysis

Variables

RR

With the video annotation tool, reviewers annotated videos a) certain breaths, b) uncertain breaths, and c) distortions on a video timeline. Based on these annotations, RR (bpm) can be calculated using each of the four methods outlined in Table 3. It was ensured that for each video, the time period considered in the denominator was the same between the five reviewers (i.e. "considered" time period). Feebris Ltd. supported Malaria Consortium in calculating RR, and distortion period (see under "**Error! Reference source not found.**") by method, reviewer and video identifier using Python, and providing the corresponding output as an Excel sheet for integration into STATA analysis. Method 4 is analogous to "bpm upper" and method 1 is analogous to "bpm lower", as per the RR automatically generated by the annotation tool (See section 'Video annotation tool' for definitions).

Table 3 Calculation of RR

Method	How the RR (bpm) is calculated	Interpretation
4	$\frac{\textit{Certain and uncertain breaths during calm}}{\textit{Total annotation time} - \textit{distortion time}} \times 60$	Less conservative WHO case management guideline
3**	$\frac{\textit{Certain breaths during calm}}{\textit{Total annotation time} - \textit{distortion time}} \times 60$	More conservative WHO case management guideline
2	$\frac{\textit{Certain and uncertain breaths}}{\textit{Total annotation time}} \times 60$	Pragmatic* WHO case management guideline = human counting with ARI timer for 60 seconds
1	$\frac{\textit{Certain breaths}}{\textit{Total annotation time}} \times 60$	Conservative pragmatic* WHO case management guideline = human counting with ARI timer for 60 seconds
Abbreviations: ARI=Acute respiratory infection, bpm=breaths per minute *Assuming that children under five are rarely fully calm and still for 60 seconds in real practice **Most appropriate reference standard		

Distortion

Distortion was defined as any time period in seconds between two annotated certain breaths that included at least one annotated distortion. Feebris Ltd. supported Malaria Consortium in calculating RR (see under "**Error! Reference source not found.**"), and distortion period by reviewer and video identifier using Python, and providing the corresponding output as an Excel sheet for integration into STATA analysis.

To account for different time period considered for analysis period between videos, standard time taken was calculated.

Standard distortion = distortion period (secs) / "considered" time period (secs)

To use the variable in analysis on video level (as opposed to observation level), average standard distortion was calculated.

Average standard distortion = Sum of distortion periods across reviewers per video/ Number of reviewers per video.

Average standard distortion was further categorized into tertiles.

Breathing status

RR (bpm) was classified into normal or fast breathing status following WHO guidelines**Error! Bookmark not defined.**.

Age group	RR range for normal breathing	RR range for fast breathing
0 to <2 months (group 1)	< 60 bpm	≥ 60 bpm
2 to <12 months (group 2)	< 50 bpm	≥ 50 bpm
12 to 59 months (group 3)	< 40 bpm	≥ 40 bpm

Time taken

Time taken was defined as the period (secs) between when the reviewer started annotating the video for phase 1 and finished annotating for phase 2. The timer was paused when the Senior Programme Officer extracted data for phase 1 and restarted once these data were extracted.

To account for different annotation period between videos, standard time taken was calculated.

Standard time taken = time taken (secs) / annotation period (secs)

To use the variable in analysis on video level (as opposed to observation level), average standard time taken was calculated.

Average standard time taken = Sum of standard time taken across reviewers per video/ Number of reviewers per video.

Average standard time taken was further categorized into tertiles.

Analysis

Data were integrated, processed and analysed in STATA 13.

All analyses were conducted for the full dataset and, where relevant, by:

- Reviewer age in years (tertiles)
- Reviewer experience in years (tertiles)
- Videos source (ARIDA, PDP)
- Country where video was taken (Ethiopia, Uganda, South Sudan)
- Gender of children in videos (Male, Female)
- Age category of children in videos (0 to < 2 months, 2 to < 12 months, 12 to 59 months)
- Average standard distortion (tertiles)

Missing data

Missing data were handled through pairwise deletion.

Descriptive

Data was described using univariate and bivariate statistics.

Agreement

The primary outcome ICC was calculated using a one-way random effects model based on the four methods for calculation of RR, respectively (see Table 3) and presented including a 95% CI

Classification by Koo & Li [13] was used to categorize ICC estimates based on the lower level of the 95% CI (Table 4).

Table 4 Interpretation of ICC estimates

ICC estimate	Interpretation
< 0.5	Poor interrater agreement
> 0.50.75	Moderate interrater agreement
> 0.75-0.90	Good interrater agreement
> 0.9	Excellent interrater agreement

The following secondary agreement measures were calculated:

- Agreement on breathing status

Agreement between five reviewers on breathing status was calculated using Fleiss kappa. A p value of < 0.05 indicates that kappa is significantly different from zero.

As per Landis and Koch [14], the strength of kappa estimates can be classified as outlined in Table 5.

Table 5 Kappa interpretation

Kappa	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

- RR range with corresponding 95% CI

RR range is calculated as the difference in RR between the reviewer with the highest RR and lowest RR per video.

- Proportion of videos with agreement for all five reviewers on breathing status (fast/normal)
- Limits of agreement

Bland Altman analysis will be conducted to extract mean difference, limits of agreement (LoA) and corresponding 95% CIs for

- Two random reviewers per video (two random out of five random reviewers of video expert panel members).
- Two same reviewers for video. As for our primary analysis each video was annotated by a random set of five reviewers, for this analysis, the two reviewers with the greatest overlap in annotated videos were selected (n=15).

Exploratory

Kruskal Wallis test with Holm's adjustment and Dunn' multiple comparison test were conducted to investigate correlation between two categorical variables age group and average distorted period (tertiles). Significance level was set at $p < 0.05$.

Qualitative analysis

Thematic analysis of the qualitative data was conducted using MAXQDA (VERBI Software, Berlin, Germany, 2016) (first FGD) and MS Excel (first and second FGD) to manage data coding, searching and retrieval. An initial coding frame for the first FGD was developed by the Epidemiologist and Senior Programme Officer, which was discussed among the team and then used to code the first transcript. AS further collated coded data into broad categories and then emerging themes. Summaries of each theme were reviewed and discussed by the research team. Upon this analysis, the team recognized the need to conduct a second FGD in order to fill some information gaps. After familiarization with the data from the second FGD, AS developed an initial coding frame and coded the second transcript accordingly. AS then reviewed the themes that emerged from the first FGD and the information that emerged from the second FGD and derived a set of consolidated themes that captured the information from both discussions. Summaries of each consolidated theme were reviewed and discussed by the research team.

Results

Study population

Video expert panel

Ten panel members were recruited for the study. Notably, the majority (80%) were male, had a mean age of 30 (95% CI: 27; 33) years and approximately 8 (95% CI: 6; 9) years' experience in their role (Table 6).

Table 6 Characteristics of video expert panel

Characteristics	No.	Column %
Number of panel members	10	100
Sex		
Male	8	80
Female	2	20
Degree		
BSC Nurse	6	60
BSC Public Health	2	20
Public Health Officer	2	20
Work place (facility type)		
Health Centre	2	20
Hospital	6	60
Health Centre & Hospital	2	20
	Mean	**[95% CI]**
Age	29.9	27.24; 32.56
Years of experience in their role	7.6	5.92; 9.28
Abbreviations: BSc=Bachelor of Science, n=number of panel members with characteristic, 95% CI=95% confidence interval		

Video sample

As per the sample size calculation during study design n=51 videos were selected from the previous ARIDA and PDP studies. One video was excluded from analysis as data could not be retrieved. Data analysis was therefore based on n=50 videos.

Over half (56%) of children in video sample were male, just over a third (38%) were in the youngest age group and the majority (78%) videos were from Ethiopia (Table 7).

Table 7 Characteristics of children in video sample

Characteristics	Total	
	No.	Column %
Total	50	100
Sex		
Male	28	56
Female	22	44
Age group of child		22
0 to < 2 months	19	38
2 to < 12 months	14	28
12 to 59 months	17	34
Country		
Ethiopia	39	78
Uganda	9	18
South Sudan	2	4

Agreement measures

Agreement is presented for RR calculated according to method 3. We consider this method to be the most appropriate reference standard due to our interpretation as a more conservative WHO case management guideline (Table 3). Results for the other methods can be found in Appendix 3 - Additional analyses.

Overall, ICC was 0.93 (95% CI: 0.89; 0.95) which is considered "good"[7] agreement on RR between five reviewers. This corresponded to a RR range of 9.19 (95% CI: 7.19; 11.19) bpm. Kappa was 0.71 which is considered substantial agreement on breathing status between five reviewers. This corresponded to all five reviewers agreeing on breathing status for 70% of the videos.

Agreement measures differed substantially between age groups: in the youngest age group (n=19), ICC was 0.87 (95% CI: 0.77; 0.94) and kappa was 0.48, whereas in the older age group (n=17), ICC was 0.94 (95% CI: 0.89; 0.98) and kappa was 0.80 (Table 8).

Table 8 Agreement stratified by age group - method 3

Categories	N	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with agreement on status
Total	50	9.19 [7.19; 11.19]	0.93 [0.89; 0.95] - good	0.71 – substantial	35/50 (70%)
0 to < 2 months	19	12.77 [9.43; 16.12]	0.87 [0.77; 0.94] - good	0.48 – moderate	9/19 (47%)
2 to < 12 months	14	9.86 [6.38; 13.35]	0.9 [0.81; 0.96] - good	0.84 – almost perfect	12/14 (86%)

[7] ICC interpretation is made using the lower limit of the 95% confidence interval

12 to 59 months	17	4.63 [2.46; 6.8]	0.94 [0.89; 0.98] - good	0.80 – substantial	14/17 (82%)

Overall, agreement between five reviewers was higher in videos with less distortion (Table 9). In exploratory analysis, videos of the age group 12 to 59 months showed a significantly lower degree of distortion than videos of the age groups 0 to < 2 months (χ^2= 2.090, p =0.04) and 2 to < 12 months (χ^2= 2.227, p =0.04).

Table 9 Agreement stratified by proportion of average distortion marked - method 3

Categories	N	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with agreement on status
Total	50	9.19 [7.19; 11.19]	0.93 [0.89; 0.95] - good	0.71 – substantial	35/50 (70%)
Highest distortion	16	15.97 [12.51; 19.42]	0.81 [0.66; 0.91] – moderate	0.57 – moderate	10/16 (63%)
Middle distortion	17	8.28 [5.89; 10.67]	0.92 [0.85; 0.97] - good	0.64 – substantial	11/17 (65%)
Lowest distortion	17	3.72 [2.75; 4.68]	0.99 [0.99; 1] - excellent	0.86 – almost perfect	14/17 (82%)

For two randomly selected reviewers per video, there was a low level of bias (mean difference) and limits of agreement (LoA) were approximately +/- 12 bpm (Table 10 and Figure 2).

Table 10 Mean difference and limits of agreement for two randomly selected reviewers, per video

Measure	Estimate [95% CI]
n	50
Mean difference [95% CI]	-0.08 [-1.96; 1.70]
Upper LoA [95% CI]	12.45 [11.21; 13.69]
Lower LoA [95% CI]	-12.60 [-13.85; -11.36]

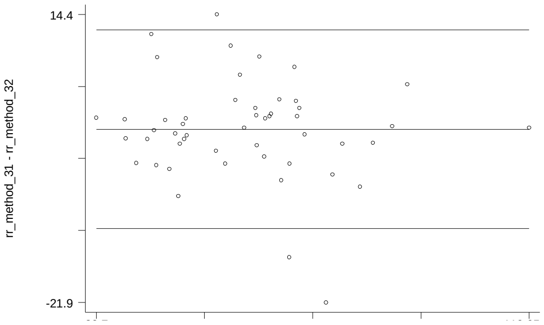


Figure 2 Bland-Altman plot for two randomly selected reviewers

For the same two reviewers per video, there was a low level of bias (mean difference) and LoA were approximately +/- 7 bpm (Table 11 and Figure 3). Note the lower sample size of n=15 videos.

Table 11 Mean difference and limits of agreement for the same reviewers, per video

Measure	Estimate [95% CI]
n	15
Mean difference [95% CI]	0.2 [-1.7; 2.1]
Lower LoA [95% CI]	-6.62 [-7.6; -5.6]
Upper LoA [95% CI]	7.02 [6.0; 8.0]

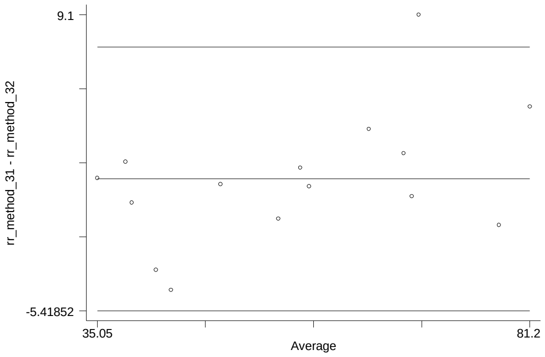


Figure 3 Bland-Altman plot for the two same reviewers

Standard time taken

On average, videos took around 30 times their length to annotate (95% CI 26.5-32.8). Videos of children in the oldest age group took the least time to annotate – 24 times their length (95% CI 20.0-27.5) (Table 12).

Table 12 Average standard time taken per video

Variable	Categories	N	Mean [95% CI]
Total	**----**	**50**	**29.7 [26.5; 32.8]**
Child age group	0 to < 2 months	19	30.6 [27.5; 33.7]
	2 to <12 months	14	35.6 [27.2; 44]
	12 to 59 months	17	23.7 [20; 27.5]
Video source	ARIDA	27	29.1 [26; 32.2]
	PDP	23	30.3 [24.4; 36.2]
Standard distorted period (average per video, tertiles)	Highest distortion	17	25.9 [21.3; 30.4]
	Middle distortion	17	33.4 [26.4; 40.4]
	Lowest distortion	16	29.7 [25.9; 33.6]

Distortion

On average, approximately one-fifth (0.19) of watched period was marked as distorted. Videos of the children in the highest age group and those from the PDP study had a lower proportion annotated as distortion (Table 13).

Table 13 Average standard distortion per video

Variable	Categories	N	Mean [95% CI]
Total	----	50	0.19 [0.15; 0.24]
Child age group	0 to < 2 months	19	0.22 [0.16; 0.29]
	2 to <12 months	14	0.24 [0.16; 0.31]
	12 to 59 months	17	0.13 [0.07; 0.19]
Video source	ARIDA	27	0.23 [0.17; 0.29]
	PDP	23	0.15 [0.1; 0.2]
Standard time taken (average per video, tertiles)	15 - 26	17	0.17 [0.04; 0.07]
	26 - 30	17	0.2 [0.15; 0.18]
	31 - 83	16	0.22 [0.34; 0.42]

Qualitative

Themes that emerged from the two FGDs were: 1) Operational factors that facilitate usability; 2) Benefits of the tool; 3) Limitations of the tool; 4) Trust in the tool. See tables below for a summary of the qualitative findings and associated quotes.

Theme 1: Video expert panel members mentioned several operational aspects that either facilitated or hindered the usability of the tool for them. Training and practice (longer than initially anticipated) were necessary, so that reviewers felt confident in annotating real videos and participating in the research. The support by trainers and the standard operating procedure document were considered very helpful while annotating videos. Reviewers appreciated the easy to use functionalities, including English as the interface language. There were suggestions that only videos with high quality should be used and functioning hardware, such as a functioning mouse, was important.

Table 14 Summary of results for theme 1

Theme	Sub-themes	Associated Quotes
1. Operational factors that facilitate usability	Keep training & practice to counter initial confusion and increase confidence	"The big thing here is practice is mandatory. Unless you exercise repeatedly, even normal breathes may confuse you, you may consider normal breathe as uncertain. It takes time to do that. But through time, as you well understand the tool, it becomes easy to count. "
	Keep standard operating procedures	"[talking about SOP] it is good if it is availed whenever we want to refer it." "How you can do something without SOPs! You do things by following it as states this is this and do this like this. You need it to refer even if you miss something. So, it was very important."
	Keep easy to use functionalities	"English is simple and easy because when you translate to Amharic, sometimes it puts us in confusion. " "We practiced each of these things including how to manipulate [i.e. using tool functionalities]. It is not difficult thing."
	Avoid using videos of insufficient quality and avoid hardware challenges	"the quality of mouse and other devices is mandatory. " "some videos were not clear to see respiration. So, it would be better if videos rerecorded by high quality HD cameras that shows clearly. "

Theme 2: Annotators mentioned two kinds of benefits the tool had in counting breaths in children. They noted that compared to counting manually, video annotation would allow them to distinguish between normal breaths, uncertain breaths, distortion and movement which would lead to a more accurate calculation of RR. They also mentioned that tool functionalities, like slowing down, changing colour adjusting brightness helped them in distinguishing these different kinds of breaths.

Table 15 Summary of results for theme 2

Theme	Sub-themes	Associated Quotes
2. Benefits of the tool	Ability to consider new elements	"this video annotation tool can identify normal breathe, uncertain breathe and distortion or other movements [...]. So, [...] the tool helps us to accurately count [...] RR of a child." "when health workers count RR, it is subjective and differs from person to person. So, the tool is to standardize it by using software"
	Tool functionalities support marking breaths in more difficult children	"By changing color, we can see whether the movement is normal breath or shallow or distortion.” "when there is distortion, we change brightness it shows movement more clearly" "When you see uncertain shallow breath, you may zoom from 1cm to 2cm and can see it.”

Theme 3: Reviewers also pointed out limitations of the tool: Videos with lots of distortions, movements and uncertain breaths were considered to be difficult to annotate with age being a factor influencing this difficulty. Reviewers found it difficult to find the exact time to mark the start/ end of distortion, to not miss shallow breaths between distortions or to identify breaths in restless and crying children as the abdomen tends to become rigid and breathing cannot be seen. While tool functionalities like slowing down or adjusting brightness helped identifying difficult breaths, using these functions remains challenging and time consuming.

Table 16 Summary of results for theme 3

Theme	Sub-themes	Associated Quotes
3. Limitations of the tool	Videos with lots of distortions, movement and uncertain breaths are difficult to annotate	"It was difficult to me to mark because you can't calm children as actual patient. And you can't seek help from other." "when child is restless and crying, the abdomen becomes rigid and breathing can't be seen." "I may miss some breathes in the mid of distortion."
	Even though functionalities help annotating "difficult" videos, using them takes time and attention	"It depends on child's age, stability and severity of disease. If child is severely sick, RR increases and marking many breathes is time consuming. I remember a video took 63 minutes from me. And spending such time on single video is a little challenging." "if there is distortion or shallow breathing, it consumes time when you go forward and backward and changing brightness etc. It can take up to 50 minutes or an hour. " "To find [respiration in mid of distortion] you go forward and backward. Because you should annotate it. It is that time challenging. " "children under two months usually breathe faster. Within this limited time, marking normal and shallow breathing requires attention."

Theme 4: Annotators showed some trust in the tool. While some suggested they would recommend this tool as a reference, some suggested they would only trust tool under certain conditions or suggested amendments. One common suggestion was that the output of the tool should only be considered if reviewers repeated their annotation and the two results agreed. Another suggestion was to exclude distortion periods from videos indicating annotating only in calm period would improve reliability of the result.

Table 17 Summary of results for theme 4

Theme	Associated quotes
4. Trust in the tool	"if difference [between result of annotation tool and other device] occurs, I will use the result of the video annotation tool because I trust my count." "Probably I can consider as good. Of course, nothing is perfect. But relatively it is good as compared to others that we have been using to count RR because it considers something that previous tools did not consider. " "To set reference, normal videos should be assessed to set reference for normal breathing. For videos with distortion, the tool should be revised [to 1) software that automatically detect distortion or 2) train experts only on distortion and test for distortion separately]" "what if distorted area is skipped from video. […] I think it is better to consider area where there is minimal distortion and it may improve accuracy. " "In the cases of differences, I may redo it up to three times with attention. When you redo repeatedly, you become confident and choose your own results. "

Discussion

Here, we present results from an interrater reliability study of a new video annotation tool for measuring RR. This is the first study assessing the reliability of RR measures derived from a manual video annotation tool allowing the annotation of certain, uncertain and distortions using five reviewers from a pool of ten.

Overall there was good agreement on RR between five reviewers as measured by ICC (0.93 [95% CI: 0.89; 0.95]) and substantial agreement on the classification of normal or fast breathing as measured by the kappa statistic (0.71; substantial agreement). Our findings show that agreement is weakest for videos of children in the lowest age group (0 to < 2 months) and for videos with the highest distortion. Similarly, in the focus group discussions, reviewers reported that the videos of young children and videos with high levels of distortions were the hardest to annotate, even when aided with additional tool functionalities. The videos of the oldest children and those with the least distortion, were also quickest to annotate.

Our results are in line with previous studies assessing interrater reliability of direct manual count or video based count to obtain RR. An interrater agreement study in north-east Tanzania measuring the agreement between two paediatricians reviewing RR videos of children aged 2-59 months found similar levels of agreement on the RR as measured by ICC (0.94 [95% CI: not reported]; excellent agreement) and on classification as measured by kappa (0.85; substantial agreement) [15]. Findings from the ARIDA study that included a subset of the videos used in this study also found similar agreement on classification between two experts counters assessing RR through manual count using the Mark 2 ARI timer (kappa = 0.83; substantial agreement) and based on counting from videos without annotation software (kappa = 0.86; substantial agreement) [11] (See Appendix 8 for poster).

Similar to our findings where agreement was lowest in videos with more distortion and younger children, authors also reported that agreement between reviewers was lower if children were agitated [15]. Obtaining reliable RR through manual counting methods (direct and based on videos) in young, agitated, moving children has been reported to be challenging [1, 16-19]. Our qualitative findings suggest that even with the tool functionalities it remained difficult for reviewers to confidently detect certain breaths during distortion, and reviewers became fatigued because some videos were time consuming, which could have affected their annotation accuracy even in calm periods of the video.

In summary, our results indicate that RR measured through the manual video annotation tool yielded a similar level of reliability to other forms of manual RR counting. Reliability of RR derived from the new video annotation tool seemed to be influenced by the same factors, with challenges in measuring RR in younger children and videos with distortion. As new aids will support the measuring of RR under real-life circumstances and account for movement automatically, a new reference should be able to provide reliable measures under the same conditions. The importance of calming the child before starting any RR count cannot be stressed enough and this should be considered when conducting any type of performance studies in future.

Historically, when training CHWs to count RR it has been suggested that an acceptable level of agreement should be +/- 2bpm, to account for a single breath variation at the start and end of the 60 second count. Similarly, the UNICEF target product profile for acute respiratory infection diagnostic aids requires new devices to measure RR within +/- 2 bpm of an established reference [20]. This would mean that any reference has to provide measures with the same reliability (or precision). Previous studies where reviewers have counted RR from a video have found that two reviewers will agree +/- 2 bpm for approximately 70% videos [11, 15]. At a WHO/UNICEF technical consultation in New York (September 2019), the scientific community acknowledged the difficulty of obtaining RR measures with this level of accuracy and suggested that Bland Altman plots with wider limits of agreement should be considered sufficiently acceptable (technical note in preparation, personal communication with Karin Kallander). Agreement for two of the same reviewers in this study, as per Bland-

Altman was comparable to the findings from the ARIDA agreement study where limits of agreement were -6.4 (95% CI: -7.5; -5.4) to 8.6 (95% CI: 7.5;9.7) (ARIDA diagnostic accuracy study report, 2017). In this study, when selecting two random reviewers per video, the limits of agreement were wider (approximately +/- 12 bpm), which is expected given that there is likely to be more human variation when randomly selecting from a larger pool of ten reviewers. There still needs to be a consensus on what the threshold for agreement on classification should be for a new aid or reference to be considered accurate or reliable, and this should be included in future target product profiles (TPPs) for all types of new RR diagnostic aids.

Agreement was also assessed using four possible methods to calculate RR. For the two methods that removed length of distorted periods, agreement as per ICC was slightly higher compared to those that included distorted periods, but it was not significant enough to change the overall interpretation of the findings and did not change the classification of the kappa statistic. Including or excluding uncertain breaths had little effect on the extent of agreement between reviewers. More work is needed to agree on the definition of a breath and this should be included in all TPPs for RR diagnostic aids going forward.

Strengths of the study were that we used a larger panel of experts than previous studies, who were randomly selected from a pool of ten experts, thus limiting the bias that may arise from a smaller pool of reviewers who might co-incidentally follow similar inaccurate patterns in counting breaths. We also used the output data about certain breaths, uncertain breaths and distortion to calculate RR in four possible ways and described how agreement between reviewers is affected when distortion and breath certainty are considered.

Limitations of the study were that some videos were of lower quality than others and the reviewers commented in the focus group discussions that for some videos it was not clear to see the child breathing. Future studies should consider allowing the reviewers to define *a priori* whether the video quality is sufficient to assess RR. Time periods considered to calculate RR for each record were edited to ensure that the same time period was considered for all five reviewers annotating the same video. Whilst the mean considered time period was 66 seconds, for 45 % (n=23) of videos, the time period considered was under 60 seconds. Future studies should ensure that the annotated period is long enough to allow for edited time periods to be longer than 60 seconds. Analysis of agreement measures was not powered for subpopulation analyses. Hence, we did not investigate if agreement outcomes were significantly different between subpopulations, and results for subpopulations may be influenced by small group sizes. Further, to classify videos into groups of low, middle and high distortion, the average annotated distortion among the five reviewers was used. Strong disagreement between reviewers on distortion within a video, e.g. very high or very low annotated distortion by a few reviewers, could have influenced this grouping of videos.

Recommendations and further work

Further analysis could be done to understand why reviewers disagree even in older, still children by overlaying the five annotated videos to see where disagreement occurs and why. These findings could support the development of a training tool to build capacity of reviewers to more accurately mark true breaths.

Further work could be done to improve the tool's usability and to make it less time consuming to annotate videos. An improved tool could be rolled out with a training package utilising high quality training videos to teach reviewers how to accurately mark breaths and distortion. Such a package could support the standardisation of reviewers to ensure they count breaths in the same way.

Updated TPPs need to be created for new types of RR diagnostic aids and reference standards which specify a definition of a breath and an agreed measure of accuracy and reliability.

Acknowledgements

This study was sponsored by the Philips Foundation. The authors would like to extend particular gratitude to the programme manager (Tedila Habte), and the video expert panel members. The study would not have been possible without the support of the Ethiopian Federal Ministry of Health and Regional Health Bureau in SNNPR.

References

1. Baker, K., et al., *Performance of Four Respiratory Rate Counters to Support Community Health Workers to Detect the Symptoms of Pneumonia in Children in Low Resource Settings: A Prospective, Multicentre, Hospital-Based, Single-Blinded, Comparative Trial.* EClinicalMedicine, 2019. **12**: p. 20-30.
2. Ward, C., et al., *Determining the Agreement Between an Automated Respiratory Rate Counter and a Reference Standard for Detecting Symptoms of Pneumonia in Children: Protocol for a Cross-Sectional Study in Ethiopia.* JMIR Res Protoc, 2020. **9**(4): p. e16531.
3. Young, M., et al., *World Health Organization/United Nations Children's Fund joint statement on integrated community case management: an equity-focused strategy to improve access to essential treatment services for children.* Am J Trop Med Hyg, 2012. **87**: p. 6-10.
4. Baker, K., et al., *Performance of Four Respiratory Rate Counters to Support Community Health Workers to Detect the Symptoms of Pneumonia in Children in Low Resource Settings: A Prospective, Multicentre, Hospital-Based, Single-Blinded, Comparative Trial.* EClinicalMedicine, 2019. **12**: p. 20-30.
5. Sinyangwe, C., et al., *Assessing the quality of care for pneumonia in integrated community case management: a cross-sectional mixed methods study.* PLoS One, 2016. **11**.
6. Black, J., et al., *Can simple mobile phone applications provide reliable counts of respiratory rates in sick infants and children? An initial evaluation of three new applications.* International Journal of Nursing Studies. **52**(5): p. 963-969.
7. Karlen, W., et al., *Improving the Accuracy and Efficiency of Respiratory Rate Measurements in Children Using Mobile Devices.* PLOS ONE, 2014. **9**(6): p. e99266.
8. World Health Organization. *Integrated management of childhood illness: chart booklet*. 2014; Available from: http://apps.who.int/iris/bitstream/10665/104772/16/9789241506823_Chartbook_eng.pdf.
9. Federal Ministry of Health Ethiopia, *Integrated Management of Newborn and Childhood Illness Assess and Classify the Sick Child Age 2 Months up to 5 Years.* 2011.
10. Philips. *ChARM Instructions for use*. 2016; Available from: http://images.philips.com/is/content/PhilipsConsumer/Campaigns/CA20160908_Global_documents/CA20160008_CO_001-AAA-en_AA-ChARM_IFU.pdf.
11. Charlotte Ward, et al., *Improving a reference standard for evaluating respiratory rate devices to diagnose symptoms of pneumonia in children under 5*. 2018: Geneva Health Forum.
12. Sekhon, M., M. Cartwright, and J.J. Francis, *Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework.* BMC Health Services Research, 2017. **17**(1): p. 88.
13. Koo, T.K. and M.Y. Li, *A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research.* Journal of chiropractic medicine, 2016. **15**(2): p. 155-163.
14. Landis, J.R. and G.G. Koch, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**(1): p. 159-74.
15. Muro, F., et al., *Variability of respiratory rate measurements in children suspected with non-severe pneumonia in north-east Tanzania.* Tropical Medicine & International Health, 2016: p. n/a-n/a.
16. Simoes, E.A., et al., *Respiratory rate: measurement of variability over time and accuracy at different counting periods.* Archives of disease in childhood, 1991. **66**(10): p. 1199-1203.

17. Amirav, I., C.K. Masumbuko, and M.T. Hawkes, *Poor Agreement and Imprecision of Respiratory Rate Measurements in Children in a Low-Income Setting.* Am J Respir Crit Care Med, 2018. **198**(11): p. 1462-1463.
18. Shah, S.A., et al., *Respiratory rate estimation during triage of children in hospitals.* J Med Eng Technol, 2015. **39**(8): p. 514-24.
19. Ginsburg, A.S., et al., *A Systematic Review of Tools to Measure Respiratory Rate in Order to Identify Childhood Pneumonia.* Am J Respir Crit Care Med, 2018. **197**(9): p. 1116-1127.
20. UNICEF. *Traget Product Profile - Acute Respiratory Infection Diagnostic Aid (ARIDA)*. 2014; Available from: https://www.unicef.org/videoaudio/PDFs/ARIDA_-_Target_Product_Profile_(2).pdf.

Appendix 3 - Additional analyses

Table 18 Children characteristics by video source

Characteristic	Categories	ARIDA		PDP		Total	
		N	Percent	N	Percent	N	Percent
Gender	Male	80	57.1	60	42.9	140	100
	Female	55	50	55	50	110	100
Age group	0 to < 2 months	45	47.4	50	52.6	95	100
	2 to <12 months	45	64.3	25	35.7	70	100
	12 to 59 months	45	52.9	40	47.1	85	100
Country	Ethiopia	135	69.2	60	30.8	195	100
	Uganda	0	0	45	100	45	100
	South Sudan	0	0	10	100	10	100

Table 19 Children characteristics by age group

Characteristic	Categories	0 to < 2 month		2 to < 12 months		12 to 59 months		Total	
		N	Percent	N	Percent	N	Percent	N	Percent
Gender	Male	45	32.1	50	35.7	45	32.1	140	100
	Female	50	45.5	20	18.2	40	36.4	110	100
Video source	ARIDA	45	33.3	45	33.3	45	33.3	135	100
	PDP	50	43.5	25	21.7	40	34.8	115	100
Country	Ethiopia	70	35.9	65	33.3	60	30.8	195	100
	Uganda	20	44.4	5	11.1	20	44.4	45	100
	South Sudan	5	50	0	0	5	50	10	100

Table 20 Average standard time taken per rating

Variable	Categories	N	Mean [95% CI]
Total	----	186	29 [26.9; 31.2]
Reviewer age group (tertiles)	26-27 years	75	27 [24.9; 29.2]
	29-30 years	53	33.4 [28; 38.8]
	31-38 years	58	27.7 [24.1; 31.2]
Reviewer experience (tertiles)	4-7 years	76	28.2 [25.8; 30.6]
	8-8.6 years	58	25.9 [22.4; 29.3]
	9-12 years	52	33.8 [28.4; 39.1]

Table 21 Agreement stratified by age group - method 1

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
Total	50	48.3 [44; 52.7]	12.19 [9.18; 15.2]	0.83 [0.76; 0.89]	0.75	37/50 (74%)

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
0 to < 2 months	19	56 [48.7; 63.2]	17.85 [12.78; 22.92]	0.75 [0.6; 0.88]	0.52	11/19 (58%)
2 to < 12 months	14	51.8 [45.3; 58.3]	12.49 [7.81; 17.16]	0.77 [0.59; 0.9]	0.76	10/14 (71%)
12 to 59 months	17	37 [31.9; 42.1]	5.62 [1.77; 9.48]	0.88 [0.78; 0.95]	0.94	16/17 (94%)

Table 22 Agreement stratified by age group - method 2

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
Total	50	49.3 [44.8; 53.7]	12.33 [9.27; 15.4]	0.83 [0.76; 0.89]	0.73	37/50 (74%)
0 to < 2 months	19	57.3 [49.9; 64.6]	17.64 [12.12; 23.16]	0.75 [0.59; 0.88]	0.55	12/19 (63%)
2 to < 12 months	14	52.9 [46.3; 59.5]	12.65 [8.18; 17.12]	0.77 [0.6; 0.91]	0.69	9/14 (64%)
12 to 59 months	17	37.4 [32.3; 42.4]	6.15 [2.2; 10.09]	0.87 [0.76; 0.94]	0.91	16/17 (94%)

Table 23 Agreement stratified by age group - method 4

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
Total	50	52.8 [47.8; 57.8]	9.28 [7.28; 11.28]	0.93 [0.9; 0.96]	0.73	36/50 (72%)
0 to < 2 months	19	62.5 [54.8; 70.2]	12.66 [9.12; 16.19]	0.87 [0.78; 0.94]	0.55	10/19 (53%)
2 to < 12 months	14	57.1 [49.3; 65]	9.68 [6.46; 12.89]	0.91 [0.82; 0.97]	0.84	12/14 (86%)
12 to 59 months	17	38.4 [33.3; 43.5]	5.18 [2.73; 7.62]	0.93 [0.87; 0.97]	0.77	14/17 (82%)

Table 24 Agreement stratified by distorted period- method 1

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
Total	50	48.3 [44; 52.7]	12.19 [9.18; 15.2]	0.83 [0.76; 0.89]	0.75	37/50 (74%)
Highest distortion	17	48.44 [38.45; 58.44]	3.54 [2.09; 4.98]	0.99 [0.98; 1]	0.9518	16/17 (94%)
Middle distortion	17	53.71 [48.44; 58.98]	11.56 [7.25; 15.87]	0.77 [0.61; 0.89]	0.8115	13/17 (76%)
Lowest distortion	16	42.55 [36.84; 48.26]	22.06 [17.43; 26.69]	0.51 [0.29; 0.75]	0.2205	8/16 (50%)

Table 25 Agreement stratified by distorted period- method 2

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
Total	50	49.3 [44.8; 53.7]	12.33 [9.27; 15.4]	0.83 [0.76; 0.89]	0.73	37/50 (74%)
Highest distortion	17	48.91 [38.69; 59.12]	3.61 [2.14; 5.08]	0.99 [0.98; 1]	0.9518	16/17 (94%)
Middle distortion	17	54.97 [49.63; 60.32]	11.33 [7.17; 15.49]	0.77 [0.61; 0.9]	0.8117	13/17 (76%)
Lowest distortion	16	43.62 [37.78; 49.46]	22.67 [17.85; 27.49]	0.51 [0.29; 0.75]	0.2115	8/16 (50%)

Table 26 Agreement stratified by distorted period- method 4

Categories	N	RR, Mean [95% CI]	RR range, Mean [95% CI]	ICC, Mean [95% CI]	Kappa for breathing status	% videos with classification agreement
Total	50	52.8 [47.8; 57.8]	9.28 [7.28; 11.28]	0.93 [0.9; 0.96]	0.73	36/50 (72%)
Highest distortion	17	49.72 [39.16; 60.29]	4.12 [3.16; 5.07]	0.99 [0.99; 1]	0.8342	14/17 (82%)
Middle distortion	17	58.92 [52.36; 65.48]	7.85 [5.79; 9.92]	0.94 [0.88; 0.97]	0.6843	12/17 (71%)
Lowest distortion	16	49.58 [41.65; 57.51]	16.28 [12.59; 19.98]	0.81 [0.66; 0.92]	0.6087	10/16 (63%)

Table 27 Average RR per video per method

RR	N	Mean [95% CI]
Method 1	50	48.3 [44; 52.7]
Method 2	50	49.3 [44.8; 53.7]
Method 3	50	52 [47.1; 56.9]
Method 4	50	52.8 [47.8; 57.8]

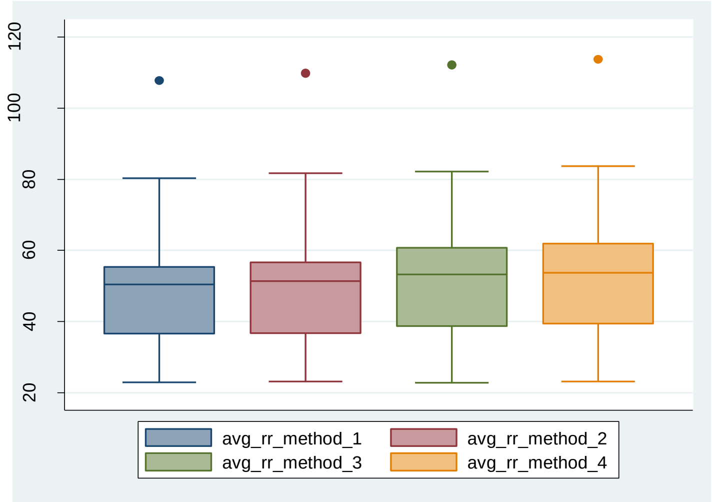


Figure 4 Box-and-whisker plot - average RR per video per method

Table 28 Length of the considered time period that was annotated by all five reviewers (seconds)

Variable	**Categories**	**N**	**Mean [95% CI]**	**Median**	**Min-Max**
Total	----	50	66.1 [61.8; 70.3]	58.6	23-108.2
Child age group	0 to < 2 months	19	66.2 [60.4; 71.9]	58.2	54.3-94.6
	2 to <12 months	14	67.9 [57; 78.7]	67	23-100.4
	12 to 59 months	17	64.5 [58.3; 70.7]	58.6	56.1-108.2
Video source	ARIDA	27	75 [69.6; 80.3]	69.5	56.2-108.2
	PDP	23	55.6 [52.5; 58.7]	57.7	23-58.8
Standard distorted period (tertiles)	0.02 - 0.10	17	59.39 [56.54; 62.23]	58.0	49.16-74.96
	0.11 - 0.23	17	62.07 [55.94; 68.21]	61.94	23-82.02
	0.23 - 0.50	16	77.42 [68.55; 86.29]	82.2	54.25-108.21

Appendix 4 - Focus group discussion guide

Focus group discussion guide for the Breath REcognition Aid To Health Experts (BREATHE) Study - Ethiopia

(Focus group discussion to be conducted by Qualitative Research Assistants in discussion with five video reviewers)

Part 1: Information and consent

Before the focus group discussion (FGD), review each individual video reviewers' signed consent form and give all participants a chance to ask questions.

Part 2: Basic information

Before the interview, qualitative research assistants should complete the following basic information for all five video reviewers:

Video reviewer (UIC)	Age	Degrees and qualifications	Number of years of experience working in a health facility	Health facility type and job title

Part 3: Focus Group Discussion

PRESS RECORD on the recorder

Say: “Focus group discussion with [UIC X, UIC Y, UIC Z etc.], by [your name], on [date]”
Say: “Thank you for agreeing to participate in this focus group discussion. The aim of this is to understand what you find easy and difficult about using the video annotation tool and to understand whether you think the tool is useful in assessing a child’s respiratory rate. During the assessment I will ask you to also provide feedback and discuss any differences in your personal experiences when using the tool. Before we start, do you have any questions?”

1. In your own words, can you describe the video annotation tool and its purpose?

2. What is your overall impression of the video annotation tool?
   - Probe: What do you like about the tool?
   - Probe: What do you not like about the tool?

3. How easy or difficult is it to use the video annotation tool?
   - Probe: How easy or difficult to:
     - Use is the English interface? Why?
     - Mark breaths? Why?
     - Mark distortion? Why?
     - Mark uncertainty? Why?
     - Move forward and backwards through the video? Why?
     - Zoom in and out? Why?
     - Speed up and slow down the video? Why?
     - Adjust the brightness of the video? Why?

4. What do you think about the length of time it takes to review one video? Why?

5. What did you think about the training for using the video annotation tool?
   - Probe: Did the training answer all of your questions? Why/why not?
   - Probe: What did you think about the length of the training? Why?
   - Probe: How helpful or unhelpful did you find the SOPs? Why?

6. How confident did you feel to use the video annotation tool after the training? Why?
7. How did your confidence in using the video annotation tool change over time? Why?

8. How useful do you think the video annotation tool is to get an accurate RR count? Why?

9. What improvements would you make to the video annotation tool? Why?

10. In future, would you recommend that the tool is used to test the accuracy of automated RR counters? Why? Why not?

Appendix 5 – information and consent form

Participant information sheet for video reviewers

Research study: Breath REcognition Aid To Health Experts (BREATHE) Study

We would like you to help with a research study. This information sheet will tell you what the research involves. Please take your time reading it. Please ask questions and you can talk it over with others if you wish.

This study is addressing the global lack of a reference standard to accurately validate new respiratory rate (RR) devices for diagnosing symptoms of pneumonia in children under the age of 5. The project will focus on building the evidence base around the accuracy of manual video annotation, as a reference standard for counting RR. The study is funded by the Philips Foundation, implemented by the Malaria Consortium and supported by the Federal Ministry of Health, Ethiopia and Regional Health Bureau in SNNPR.

As a consultant video reviewer in this study, you will review and annotate videos of children breathing using an annotation tool. You might also be asked to take part in a Focus Group Discussion (FGD) at the end of data collection to get an insight into how easy it is to use the tool and whether you think it is acceptable.

Why have I been chosen for the study?

You have been chosen for this study because you are a health professional with experience counting RR in children and are proficient using IT software. You will use the annotation tool to mark breaths on videos of children to obtain a RR. After this data collection, you may be asked to give your opinion on the annotation tool through FGD, where you will be asked questions about how easy it is to use the tool and whether you think it is acceptable.

What happens if I agree to take part?

You will review and annotate up to 30 videos over a period of 3 to 4 weeks as part of your video consultancy and then may be asked to participate in a FGD. Taking part in the FGD is completely your choice. If you decide not to participate, there will be no negative consequences. If you decide to take part, you will be asked to give consent by signing a form. The FGD will take a maximum of three hours. **You can choose to withdraw from the FGD at any point during the FGD or up until the completion of data analysis. Data analysis is expected to be concluded by *end of May.*** There will be no negative consequences should you decide to withdraw your consent. If you would like to withdraw from the FGD please inform the Project Manager either in person or on the contact number listed at the end of this form.

What are the benefits of taking part?
There are no direct benefits to the video reviewers for participating in the research study.

What are the possible disadvantages and risks of taking part?
There are no disadvantages or risks to video reviewers for taking part in the research study.

How is my information being recorded?
Your answers for the video annotation will be collected and analysed.
Your answers for the FGD will be audio recorded and then transcribed in writing into Amharic and then English.

Will my participation in the study be kept confidential?
You will not be identified or identifiable by name in any reports of publications.

What will happen to the results of the study?
They will be used to improve pneumonia diagnosis in children under 5. The results will also be published in medical journals. The data collected might also be used by Malaria Consortium in future research studies, if approved by a relevant ethics committee. Furthermore, anonymised data collected may be made publicly available at the end of this study and may be used for purposes not related to this study. However, it will not be possible to identify you from this data.

What happens if the research study stops earlier than expected?
If it does, we will provide you with clear information as to why.

Who is doing and paying for the research?
Malaria Consortium is implementing the study, with financial support from the Philips Foundation.

Will I be compensated?
The FGD will be compensated at the same rate as the rest of your consultancy for this research project.

If you have any questions at any time please ask a member of the research team or contact:

Tedila Habte

Malaria Consortium, Hawassa

t.habte@malariaconsortium.org

0462204415

If you have serious concerns regarding the conduct of this study and would prefer to report these to an independent body, contact the ethics review board that oversees this study:

Contact: W/ro. Emebet Mekonen - Research and Technology Transfer Process owner

Address: Hawassa, SNNPR Health Bureau

Email: emekonnen62@yahoo.com

Phone number: 0911817744

Once you have read and fully understood this information, please consider whether you would like to participate. If you agree to participate in this research study, please sign the 'Participant consent form'.

Thank you for your time.

Participant consent form

1. I confirm that I have read and understood the information sheet dated March 2019, explaining the above research study and I have had the opportunity to ask questions about the study.
2. I confirm that I am 18 years or older.
3. I understand that I will participate in the RR reference standard study that uses videos collected in a previous study conducted by Malaria Consortium,
4. I understand that I may be asked to participate in a Focus Group Discussion (FGD) at the end of data collection which will be audio recorded. I understand that I am free to decline participation or can decline answering certain questions.
5. I understand that my name will not be linked to the research materials and any personal information that could identify me will be kept strictly confidential.
6. I understand that my responses will be anonymised and that I will not be identified or identifiable in any report, publications, or presentations that result from my participation in this study.
7. I agree for the data collected during the video annotation and the FGD to be used in future research by Malaria Consortium and that after the study, my anonymised data may be made publicly available and may be used for purposes not related to this study. However, it will not be possible to identify me from this data.
8. I agree to take part in the above research study.

Name of participant	**Date:**	**Signature:**
Name of person taking consent:*	**Date:**	**Signature:**

*To be signed and dated in the presence of the participant.

UIC:

Contact number of Malaria Consortium's Project Manager: 0462204415

Appendix 6 – Focus Group Discussion topic guide

Focus group discussion guide for the Breath REcognition Aid To Health Experts (BREATHE) Study - Ethiopia

(Focus group discussion to be conducted by Qualitative Research Assistants in discussion with five video reviewers)

Part 1: Information and consent

Before the focus group discussion (FGD), review each individual video reviewers' signed consent form and give all participants a chance to ask questions.

Part 2: Basic information

Before the interview, qualitative research assistants should complete the following basic information for all five video reviewers:

Video reviewer (UIC)	Age	Degrees and qualifications	Number of years of experience working in a health facility	Health facility type and job title

Part 3: Focus Group Discussion

PRESS RECORD on the recorder

Say: "Focus group discussion with [UIC X, UIC Y, UIC Z etc.], by [your name], on [date]"

Say: "Thank you for agreeing to participate in this focus group discussion. The aim of this is to understand what you find easy and difficult about using the video annotation tool and to understand whether you think the tool is useful in assessing a child's respiratory rate (RR). During the assessment I will ask you to also provide feedback and discuss any differences in your personal experiences when using the tool. Before we start, do you have any questions?"

1. In your own words, can you describe the video annotation tool and its purpose?

2. What is your overall impression of the video annotation tool?
   - ➢ Probe: What do you like about the tool?
   - ➢ Probe: What do you not like about the tool?

3. How easy or difficult is it to use the video annotation tool?
   - ➢ Probe: How easy or difficult to:
     - o Use is the English interface? Why?
     - o Mark breaths? Why?
     - o Mark distortion? Why?
     - o Mark uncertainty? Why?
     - o Move forward and backwards through the video? Why?
     - o Zoom in and out? Why?
     - o Speed up and slow down the video? Why?
     - o Adjust the brightness of the video? Why?

4. What do you think about the length of time it takes to review one video? Why?

5. What did you think about the training for using the video annotation tool?
   - ➢ Probe: Did the training answer all of your questions? Why/why not?
   - ➢ Probe: What did you think about the length of the training? Why?
   - ➢ Probe: How helpful or unhelpful did you find the SOPs? Why?

6. How confident did you feel to use the video annotation tool after the training? Why?

7. How did your confidence in using the video annotation tool change over time? Why?

8. How useful do you think the video annotation tool is to get an accurate RR count? Why?

9. What improvements would you make to the video annotation tool? Why?

10. In future, would you recommend that the tool is used to test the accuracy of automated RR counters? Why? Why not?

Appendix 7 – Consent form for ARIDA Diagnostic accuracy project

Acute Respiratory Infection Diagnostic Aids (ARIDA) Field Trials Controlled Accuracy Study, Ethiopia

Form 2a: Parent or guardian of child under five information sheet for consent

We would like you to help with a research study. This information sheet will tell you what the research involves. Please take your time reading it. It can be read out to you if you choose. Please ask questions and you can talk it over with others if you wish.

Malaria Consortium, in partnership with UNICEF and the Federal Ministry of Health, Ethiopia, are conducting Acute Respiratory Infection Diagnostic Aids (ARIDA) Field Trials in Ethiopia.

The aim of the ARIDA field trials is to evaluate automated RR counting aids for use by health extension workers (HEWs) and frontline health facility workers (FLHFWs) in detection of fast breathing pneumonia, depending on the age of your child. This stage of the research will determine the performance of the ARIDA test device in children under 5 in a controlled setting using two evaluations: accuracy and repeatability. A third evaluation will measure RR fluctuation over time after ARIDA test device attachment, in normal breathing children aged 2-59 months, in a controlled setting. Performance of expert clinicians using a standard RR timer will also measured simultaneously.

Why have I been chosen for the study? You are the parent or guardian of a child under 5, and you have brought your child to hospital for diagnosis and treatment. Approximately 300 children under 5 will participate in this stage of the study, which will enable us to evaluate the ARIDA test device to diagnose pneumonia in a controlled setting.

What happens if I agree to take part? Your child will be involved in up to two RR evaluations (accuracy, repeatability, fluctuation) lasting up to 30 minutes in total. Whether you take part is your choice. Participation is completely voluntary; you may choose not to take part or to stop at any time. You will continue to receive the same medicines as usual if you do or do not agree to participate. If you participate you will help us find out the best way to diagnose pneumonia in the community.

The evaluations will focus on recording RR to diagnose pneumonia in children. A full assessment of your child's current illness (if any) and a full assessment of signs and symptoms of pneumonia will be obtained by the research nurse. Before any RR evaluations take place, the research team will ensure your child is calm and comfortable.

For the accuracy evaluation, an ARIDA test device will be used to count the number of breaths in one minute and an expert clinician will conduct a manual RR count at the same time. For the repeatability evaluation, two ARIDA test devices will be used to count the number of breaths in one minute and two expert clinicians will conduct manual RR counts at the same time. A video will also be taken for both of these evaluations. There will be up to three attempts to obtain a RR reading for both of these evaluations.

For the RR fluctuation evaluation, an ARIDA test device will be attached to your child for approximately 6 minutes and an expert clinician will conduct three manual counts during this period. No video recording will be taken.

What are the benefits of taking part? There are no direct benefits to you or your child, but this study hopes to improve the care of all children with pneumonia in the future.

What happens during the assessment? Depending on the age and RR of your child, they will participate in up to two of the RR evaluations (accuracy, repeatability, fluctuation).

What are the possible disadvantages and risks of taking part?

There are no added risks involved in participating in this study. The measures that will be used to observe RR in your child will not be able to penetrate your child's body. A blood sample or sample of other body fluids from your child will not be taken. As the device is placed on your child's skin it will not cause pain or injury. Due to the time taken to conduct the evaluation(s), your child may experience minor discomfort. Whether you agree or not to take part you will continue to receive the same medicines and tests as usual.

Will my participation in the study be kept confidential? Yes. Your child's name will be stored by number in the research forms, and this will be available only to the research team working on the study.

How is the data being recorded? Your child's data will be recorded on tablets. Data will also be recorded on video (for accuracy and repeatability evaluations only).

What will happen to the results of the study? They will be used to improve pneumonia diagnosis in children under 5. The results will also be published in medical journals. You and your child will not be identified or identifiable by name in any reports of publications.

What happens if the research study stops earlier than expected? If it does, we will provide you with clear information as to why.

Who is doing and paying for the research? Malaria Consortium is carrying out the study, with support from "La Caixa Foundation" and in partnership with UNICEF.

Who else supports this research? This research is supported by Federal Ministry of Health in Addis Ababa.

If you have any questions at any time, please ask a member of the research team or you can contact:

Tedila Habte, Project Officer, Malaria Consortium Ethiopia, Phone: +251931404500

Dr Geremew Tarekegne Tsegaye, AHRI/ALERT Ethics Review Committee (AAERC), Phone: +251910809415

Once you have read and fully understood this information, please consider whether you would like your child to participate and proceed to completing sheet 2b 'consent form' if you wish to give consent for your child's participation.

Thank you for your time.

Acute Respiratory Infection Diagnostic Aids (ARIDA) Field Trials Controlled Accuracy Study, Ethiopia

Form 2b: Parent or guardian of child under five consent form

Research study: ARIDA Controlled Accuracy Study

1. I confirm that I have read and understood the information sheet dated _____________________, explaining the above research project and I have had the opportunity to ask questions about the study.

2. I am 16 years or older.

3. I understand that giving consent for my child's participation is voluntary and that I am free to withdraw my consent at any time without giving any reason and without any negative consequences. In addition, should I not wish to answer any particular questions, I am free to decline.

4. I understand that my name and that of my child will not be linked to the research materials and any personal information that could identify me or my child will be kept strictly confidential. I understand that my responses will be anonymised and that I or my child will not be identified or identifiable in any report, publications or presentations that result from this research.

5. I agree for the data collected from my child to be used in future research.

6. I give permission for this evaluation to be video recorded, to be used only for analysis.

7. I agree to take part in the above research project.

*To be signed and dated in the presence of the participant

Name of parent or guardian:

Date:

Signature/thumb print:

If parent or guardian is illiterate please obtain the additional signature of a witness:

Name of witness:

Date:

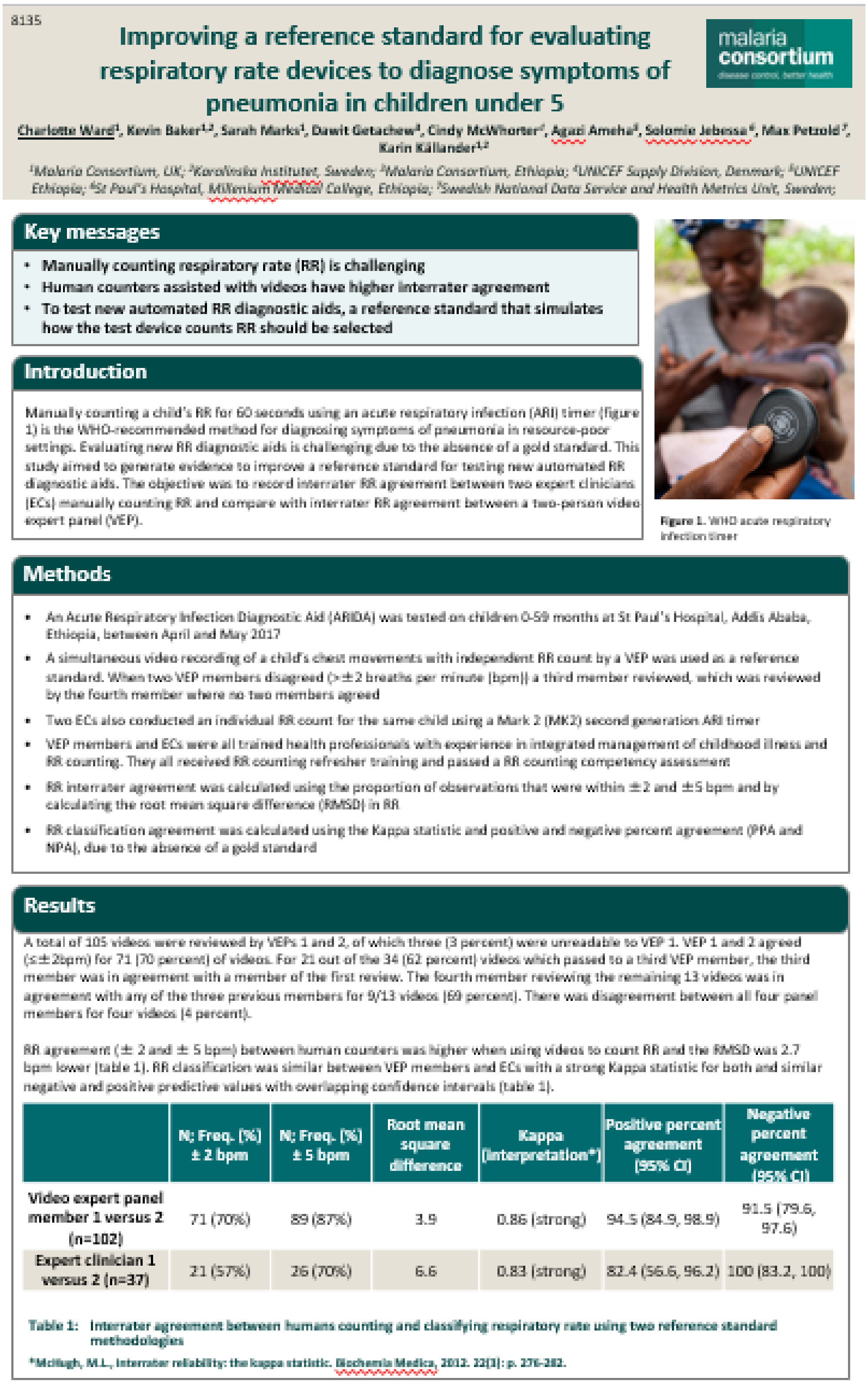
Signature/thumb print:

Name of person taking consent:

Date:

Signature/thumb print:

Appendix 8 – Geneva Health Forum poster presentation 2017



8135

Improving a reference standard for evaluating respiratory rate devices to diagnose symptoms of pneumonia in children under 5

Charlotte Ward[1], Kevin Baker[1,2], Sarah Marks[1], Dawit Getachew[3], Cindy McWhorter[4], Agazi Ameha[3], Solomie Jebessa[6], Max Petzold[7], Karin Källander[1,2]

[1]Malaria Consortium, UK; [2]Karolinska Institutet, Sweden; [3]Malaria Consortium, Ethiopia; [4]UNICEF Supply Division, Denmark; [5]UNICEF Ethiopia; [6]St Paul's Hospital, Millenium Medical College, Ethiopia; [7]Swedish National Data Service and Health Metrics Unit, Sweden;

Key messages

- **Manually counting respiratory rate (RR) is challenging**
- **Human counters assisted with videos have higher interrater agreement**
- **To test new automated RR diagnostic aids, a reference standard that simulates how the test device counts RR should be selected**

Introduction

Manually counting a child's RR for 60 seconds using an acute respiratory infection (ARI) timer (figure 1) is the WHO-recommended method for diagnosing symptoms of pneumonia in resource-poor settings. Evaluating new RR diagnostic aids is challenging due to the absence of a gold standard. This study aimed to generate evidence to improve a reference standard for testing new automated RR diagnostic aids. The objective was to record interrater RR agreement between two expert clinicians (ECs) manually counting RR and compare with interrater RR agreement between a two-person video expert panel (VEP).

Figure 1. WHO acute respiratory infection timer

Methods

- An Acute Respiratory Infection Diagnostic Aid (ARIDA) was tested on children 0-59 months at St Paul's Hospital, Addis Ababa, Ethiopia, between April and May 2017
- A simultaneous video recording of a child's chest movements with independent RR count by a VEP was used as a reference standard. When two VEP members disagreed (>±2 breaths per minute (bpm)) a third member reviewed, which was reviewed by the fourth member where no two members agreed
- Two ECs also conducted an individual RR count for the same child using a Mark 2 (MK2) second generation ARI timer
- VEP members and ECs were all trained health professionals with experience in integrated management of childhood illness and RR counting. They all received RR counting refresher training and passed a RR counting competency assessment
- RR interrater agreement was calculated using the proportion of observations that were within ±2 and ±5 bpm and by calculating the root mean square difference (RMSD) in RR
- RR classification agreement was calculated using the Kappa statistic and positive and negative percent agreement (PPA and NPA), due to the absence of a gold standard

Results

A total of 105 videos were reviewed by VEPs 1 and 2, of which three (3 percent) were unreadable to VEP 1. VEP 1 and 2 agreed (≤±2bpm) for 71 (70 percent) of videos. For 21 out of the 34 (62 percent) videos which passed to a third VEP member, the third member was in agreement with a member of the first review. The fourth member reviewing the remaining 13 videos was in agreement with any of the three previous members for 9/13 videos (69 percent). There was disagreement between all four panel members for four videos (4 percent).

RR agreement (± 2 and ± 5 bpm) between human counters was higher when using videos to count RR and the RMSD was 2.7 bpm lower (table 1). RR classification was similar between VEP members and ECs with a strong Kappa statistic for both and similar negative and positive predictive values with overlapping confidence intervals (table 1).

	N; Freq. (%) ± 2 bpm	N; Freq. (%) ± 5 bpm	Root mean square difference	Kappa (Interpretation*)	Positive percent agreement (95% CI)	Negative percent agreement (95% CI)
Video expert panel member 1 versus 2 (n=102)	71 (70%)	89 (87%)	3.9	0.86 (strong)	94.5 (84.9, 98.9)	91.5 (79.6, 97.6)
Expert clinician 1 versus 2 (n=37)	21 (57%)	26 (70%)	6.6	0.83 (strong)	82.4 (56.6, 96.2)	100 (83.2, 100)

Table 1: Interrater agreement between humans counting and classifying respiratory rate using two reference standard methodologies

*McHugh, M.L., Interrater reliability: the kappa statistic. Biochemia Medica, 2012. 22(3): p. 276-282.