RESEARCH ARTICLE

REVISED Identification of thresholds for accuracy comparisons of heart rate and respiratory rate in neonates [version 2; peer review: 2 approved, 1 approved with reservations, 1 not approved]

Jesse Coleman[1], Amy Sarah Ginsburg[2], William M. Macharia[3], Roseline Ochieng[3], Guohai Zhou[4], Dustin Dunsmuir[5], Walter Karlen[6], J. Mark Ansermino[5]

[1]Evaluation of Technologies for Neonates in Africa (ETNA), Aga Khan University Hospital, Nairobi, Kenya
[2]University of Washington, Seattle, WA, 98195, USA
[3]Department of Paediatrics, Aga Khan University Hospital, Nairobi, Kenya
[4]Center for Clinical Investigation, Brigham and Women's Hospital, Boston, MA, 02115, USA
[5]Anesthesiology, Pharmacology & Therapeutics, The University of British Columbia, Vancouver, BC, V6T 1Z3, Canada
[6]Mobile Health Systems Lab, Department of Health Sciences and Technology, ETH Zürich, Zürich, 8092, Switzerland

v2 **First published:** 10 Jun 2021, **5**:93
https://doi.org/10.12688/gatesopenres.13237.1
Latest published: 08 Oct 2021, **5**:93
https://doi.org/10.12688/gatesopenres.13237.2

Abstract

Background: Heart rate (HR) and respiratory rate (RR) can be challenging to measure accurately and reliably in neonates. The introduction of innovative, non-invasive measurement technologies suitable for resource-constrained settings is limited by the lack of appropriate clinical thresholds for accuracy comparison studies.
Methods: We collected measurements of photoplethysmography-recorded HR and capnography-recorded exhaled carbon dioxide across multiple 60-second epochs (observations) in enrolled neonates admitted to the neonatal care unit at Aga Khan University Hospital in Nairobi, Kenya. Trained study nurses manually recorded HR, and the study team manually counted individual breaths from capnograms. For comparison, HR and RR also were measured using an automated signal detection algorithm. Clinical measurements were analyzed for repeatability.
Results: A total of 297 epochs across 35 neonates were recorded. Manual HR showed a bias of -2.4 (-1.8%) and a spread between the 95% limits of agreement (LOA) of 40.3 (29.6%) compared to the algorithm-derived median HR. Manual RR showed a bias of -3.2 (-6.6%) and a spread between the 95% LOA of 17.9 (37.3%) compared to the algorithm-derived median RR, and a bias of -0.5 (1.1%) and a spread between the 95% LOA of 4.4 (9.1%) compared to the algorithm-derived RR count. Manual HR and RR showed repeatability of 0.6 (interquartile

Open Peer Review

Reviewer Status ? ✓ ✗ ✓

	Invited Reviewers			
	1	**2**	**3**	**4**
version 2 (revision) 08 Oct 2021	? report	✓ report	✗ report	✓ report
version 1 10 Jun 2021	✗ report			

1. **Gordon B. Drummond**, University of Edinburgh, Edinburgh, UK
2. **Kevin Baker**, Malaria Consortium, London, UK
Karolinska Institute, Stockholm, Sweden
3. **Robert E. Kearney**, McGill University, Montreal, Canada

range (IQR) 0.5-0.7), and 0.7 (IQR 0.5-0.8), respectively.
Conclusions: Appropriate clinical thresholds should be selected *a priori* when performing accuracy comparisons for HR and RR. Automated measurement technologies typically use a smoothing or averaging filter, which significantly impacts accuracy. A wider spread between the LOA, as much as 30%, should be considered to account for the observed physiological nuances and within- and between-neonate variability and different averaging methods. Wider adoption of thresholds by data standards organizations and technology developers and manufacturers will increase the robustness of clinical comparison studies.

Keywords
neonatal vital sign measurement, monitoring, heart rate, respiratory rate, accuracy, validation

4. **AbdelKebir Sabil**, Cloud Sleep Lab, Paris, France

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Jesse Coleman (denots@gmail.com)

Author roles: Coleman J: Investigation, Methodology, Project Administration, Software, Visualization, Writing – Original Draft Preparation; **Ginsburg AS**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Macharia WM**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Ochieng R**: Methodology, Resources, Supervision, Writing – Review & Editing; **Zhou G**: Conceptualization, Formal Analysis, Methodology, Software, Supervision, Validation, Visualization, Writing – Review & Editing; **Dunsmuir D**: Formal Analysis, Resources, Software, Writing – Review & Editing; **Karlen W**: Methodology, Resources, Software, Writing – Review & Editing; **Ansermino JM**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Bill & Melinda Gates Foundation [OPP1196617]
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Coleman J, Ginsburg AS, Macharia WM *et al.* **Identification of thresholds for accuracy comparisons of heart rate and respiratory rate in neonates [version 2; peer review: 2 approved, 1 approved with reservations, 1 not approved]** Gates Open Research 2021, **5**:93 https://doi.org/10.12688/gatesopenres.13237.2

First published: 10 Jun 2021, **5**:93 https://doi.org/10.12688/gatesopenres.13237.1

REVISED **Amendments from Version 1**

Based on helpful feedback from external reviewers, we have updated our manuscript to clarify the methods we used to synchronize the heart rate and respiratory rate data, along with the aims of the study and an updated Figure 2 to include 95% confidence intervals for the upper and lower limits of agreement.

Any further responses from the reviewers can be found at the end of the article

Introduction

There is a high risk of mortality during the neonatal period, particularly in resource-constrained settings[1]. Continuous monitoring of neonatal vital signs enables early detection of physiological deterioration and potential opportunities for life-saving interventions[2–4]. The development of new, innovative, non-invasive, multiparameter continuous physiological monitors specifically for neonates offers the promise of improving clinical outcomes in this vulnerable population. However, before use, these technologies should be tested in real-world situations to determine accuracy and clinical feasibility.

A neonate's marked physiological variability, small size, and often fragile condition can offer challenges when measuring and monitoring vital signs. A lack of neonatal clinical validation standards further undermines the development of continuous monitors clinically validated specifically for neonates. Determining the accuracy of new continuous monitors is an essential step in bringing these technologies to market[5,6].

The Evaluation of Technologies for Neonates in Africa (ETNA) platform aims to independently establish the accuracy and feasibility of novel continuous monitors suitable for use in neonates in resource-constrained settings[7]. To determine accuracy and agreement, new technologies are compared against existing reference methods or technologies[8]. Before the comparison process can proceed, a clinical reference verification step is necessary to determine appropriate accuracy thresholds[7]. These *a priori* thresholds determine the target level of agreement required and thus, the success or failure of an investigational technology. This study describes the verification processes we conducted with a clinical reference technology in order to determine appropriate heart rate (HR) and respiratory rate (RR) accuracy thresholds to use in subsequent new continuous monitors accuracy comparisons.

Methods

Study design

This was a cross-sectional study which aimed to identify the natural variation in neonatal HR and RR in order to identify appropriate accuracy thresholds for use in an accuracy comparison of continuous monitors.

Setting and participants

Study participants were neonates admitted for observation and care in the maternity ward, neonatal intensive care, and the neonatal high dependency units at Aga Khan University Hospital in Nairobi, Kenya (AKUHN). Between June and August 2019, caregivers were approached, recruited, and sequentially screened for enrolment by trained study staff during routine newborn intake procedures. To minimize potential selection bias, all caregivers were approached in a sequential manner, as much as possible and introduced to the study using a standardized recruitment script. Final eligibility determination was dependent on medical history results, physical examination, an appropriate understanding of the study by the caregiver, and completion of the written informed consent process (Table 1).

Study procedures

The Masimo Rad-97 Pulse CO-Oximeter® with NomoLine Capnography (Masimo Corporation, Irvine, CA, USA) was selected as the reference technology based on validated oxygen saturation (SpO_2) accuracy measurement in neonates[9–11]. During study participation, trained and experienced study nurses attached the Rad-97 to neonates and conducted manual HR measurements (counting over 60-second epochs) every 10 minutes for the first hour and once per hour of participation thereafter, following World Health Organization (WHO) guidance for HR measurement in neonates[12]. Photoplethysmographic HR was also measured via the Masimo Rad-97 pulse oximetry skin sensor attached to the neonate's foot. RR was measured by capnography using an infant/pediatric nasal cannula to collect the neonate's exhaled carbon dioxide (CO_2) levels. Duration of data collection length was set at a minimum of one hour, with no upper limit. Neonates exited from the study upon discharge from the ward or by caregiver request.

Data collection and analysis

Using a custom Android (Google, Mountain View, CA, USA) application, raw data was collected from the Masimo Rad-97 in real-time through a universal serial bus (USB) asynchronous connection and parsed in C (Dennis Ritchie & Bell Labs, USA). Instantaneous HR was obtained from the timing of the pulse oximetry signal quality index (PO-SQI). The plethysmogram waveform was sampled at 62.5 Hz with the PO-SQI identified by the Masimo Rad-97 at the peak of each heartbeat. The CO_2 waveform was sampled at approximately 20 Hz from the capnography channel. The parsed output included an accurate time stamp for each entry in the waveform data output to facilitate synchronization and analysis. Data were recorded and stored on a secure AKUHN-hosted REDCap server[13].

We analyzed the CO_2 waveform data using a breath detection algorithm developed in MATLAB (Math Works, USA) and based on adaptive pulse segmentation[14]. In addition to providing a RR, the algorithm analyzed the waveform's shape and identified the breath duration (waveform trough to trough) for each breath. From the breath duration, we calculated a RR based on the median breath duration within the epoch. We developed a custom capnography quality score (CO_2-SQI) based on capnography features to assist with data selection. HR and RR counts and medians, along with signal quality

Table 1. Study eligibility criteria and definitions.

Eligibility criteria	
Inclusion criteria	• Male or female neonate, corrected age of <28 days • Willingness and ability of neonate's caregiver to provide informed consent and to be available for follow-up for the planned duration of the study
Exclusion criteria	• Receiving mechanical ventilation or continuous positive airway pressure • Skin abnormalities in the nasopharynx and/or oropharynx • Contraindication to the application of skin sensors • Known arrhythmia • Any medical or psychosocial condition or circumstance that, in the opinion of the investigators, would interfere with the conduct of the study or for which study participation might jeopardize the neonate's health
Study definitions	
Epoch	A 60-second period of time
Heartbeat	One pulsation of the heart, including one complete contraction and dilatation
Heart rate (HR)	Number of heart beats within an epoch
Breath	One cycle of inhalation and exhalation
Breath duration	Length of time from the start to the end of a single breath
Respiratory rate (RR)	Number of breaths initiated within an epoch
Pulse oximetry signal quality index (PO-SQI)	Automated indicator of signal quality from the plethysmographic recording.
CO_2-SQI	Algorithm-defined indicator of signal quality from the capnography channel
Accuracy	The closeness a measured value is from the true value
Repeatability	The closeness of the results of successive measurements of the same measure
Agreement (between measures)	The consistency between two sets of measurements
Accuracy Threshold	A pre-specified value used to determine if a set of measurements has achieved a sufficient accuracy when compared with a reference value
Precision	The closeness of measurements to each other

metrics from the MATLAB signal detection algorithm, were analyzed using R version 4.0.3[15]. Capnogram waveforms were generated with two seconds added at the beginning and end of each epoch to facilitate manual breath counting within the epoch.

To ensure temporal alignment between measurements, HR and RR epochs were synchronized across source data devices. For HR, alignment was done using a timestamp in REDCap that was set by the study nurse as HR counting was initiated. Before analysis, this timestamp was synchronized with the same timestamp in the custom Android application. Both the REDCap and Android servers were connected via the internet to a Network Time Protocol (NTP) server. Alignment of RR epochs was based on the Android application timestamp. All RR waveforms were compared visually to further ensure epoch synchronization.

One of the authors (JMA, a pediatric anesthesiologist) reviewed the capnogram tracings and discarded plots with marked variability or a significant duration of an artifact that would have made breaths challenging to count. The remaining plots were provided to two trained observers to independently count all breaths within each epoch using a set of predefined rules created by the investigators (Table 2). The two independent counts were averaged, and if the number of breaths counted by the two observers varied by more than three breaths per epoch, a third trained observer independently counted the plot, and the two closest counts were averaged.

Measurement repeatability was estimated using linear mixed-effects models based on the between- and within-neonate variability for each data source using R version 4.0.3[16]. Agreement between data collection methods was assessed using the method described by Bland-Altman for replicated observations and

reported as a mean bias with 95% confidence intervals (CIs), 95% upper and lower limits of agreement (LOA), and as a root mean square deviation (RMSD)[17]. The aim was to identify practical threshold limits using data from the clinical reference technology verification process.

Sample size

We estimated that 20 neonates with ten replications each would give a 95% CI LOA between two methods of +/-0.76 times the standard deviation (SD) of their differences. Sample size estimates for method comparison studies typically depend on the CI required around the LOA, and sample sizes of 100 to 200 provide tight CIs[17]. We aimed for a sample size of at least 30 neonates to ensure a diverse population and sufficient replications for tight CIs.

Ethical approval

The study was conducted per the International Conference on Harmonisation Good Clinical Practice and the Declaration of Helsinki 2008. The protocol and other relevant study documents were approved by Western Institutional Review Board (20191102; Puyallup, Washington, USA), Aga Khan University Nairobi Research Ethics Committee (2019/REC-02 v2; Nairobi, Kenya), Kenyan Pharmacy and Poisons Board (19/05/02/2019(078)) and Kenyan National Commission for Science, Technology and Innovation (NACOSTI/P/19/68024/30253). Written informed consent was obtained in English or Swahili by trained study staff from each neonate's caregiver according to a checklist that included ascertainment of caregiver comprehension.

Results

Between June and August 2019, 35 neonates were enrolled, and 297 clinical observations were completed with a mean of 8.4 (SD 1.7) observations per neonate (Table 3; Figure 1) and a median data collection time of 4 hours, 5 minutes (interquartile range (IQR) 3:52-4:45)[18]. The manual HR measurements were found to have a non-normal distribution with skewness of 0.76 and kurtosis of 3.60 ($p<0.001$). The median manual HR measurement for all observations was 134 (IQR 126-143) beats per minute (bpm).

The manual HR demonstrated a negative bias of -2.4 (-1.8%) compared to the median PO-SQI HR, and a marked spread between the 95% LOA of 40.3 (29.6%). The RMSD was 10.5 (7.7%). Removing data from a single outlier neonate resulted in a smaller bias of -1.4 (-1.0%), a tighter spread between the 95% LOA of 24.7 (18.2%), and a lower RMSD of 6.4 (4.7%) (Table 4; Figure 2).

Moderate repeatability was demonstrated with approximately 62% (95% CI 47%-73%) of the manual HR variability being due to differences between neonates (Table 5, Figure 3A). Since the 95% CI for manual HR crossed 50%, the between- and within-neonate variability appeared to be comparable, with neither causing significantly more variability than the other.

Manual RR from capnograms were found to have a non-normal distribution with skewness of 0.61 and kurtosis of 2.96 ($p=0.027$). The median manual RR measurement for all observations was 47 (IQR 39-56) breaths per minute. The manual RR compared to the algorithm-derived median RR showed a negative bias of -3.2 (-6.6%) and a marked spread between the 95% LOA of 17.9 (37.3%). The RMSD was 5.5 (11.4%). Comparing the manual RR to the algorithm-derived RR count showed a smaller bias of -0.5 (-1.1%) and a tighter spread between the 95% LOA of 4.4 (9.1%). The RMSD was 1.2 (2.5%).

Table 2. Rules for identifying breaths based on graphical waveform plots.

1. Count peaks of the waveform that are within the white background. Ignore peaks that are within the grey background on either side of the image.
2. A peak should be counted as a breath when the peak of the waveform is above 15mmHg, the lower horizontal blue line.
3. If the peak does not reach the lower horizontal blue line at 15 mmHg, to be counted as a breath, the peak should reach at least 50% of the mean peak.
4. The waveform should dip down to the normal baseline (either below 15 mmHg, the lower horizontal blue line, or based on other breaths). If the waveform does not reach below this point, then this is considered part of the same (double) peak and only counted as a breath once.

Table 3. Neonate demographic data.

Sex			Age at participation (days)		Gestation at birth (weeks)		Weight at birth (grams)	
Female	Male	Other	Median	IQR	Median	Range	Median	IQR
22	13	0	2	0-4	33	32-34	1500	1260-1600

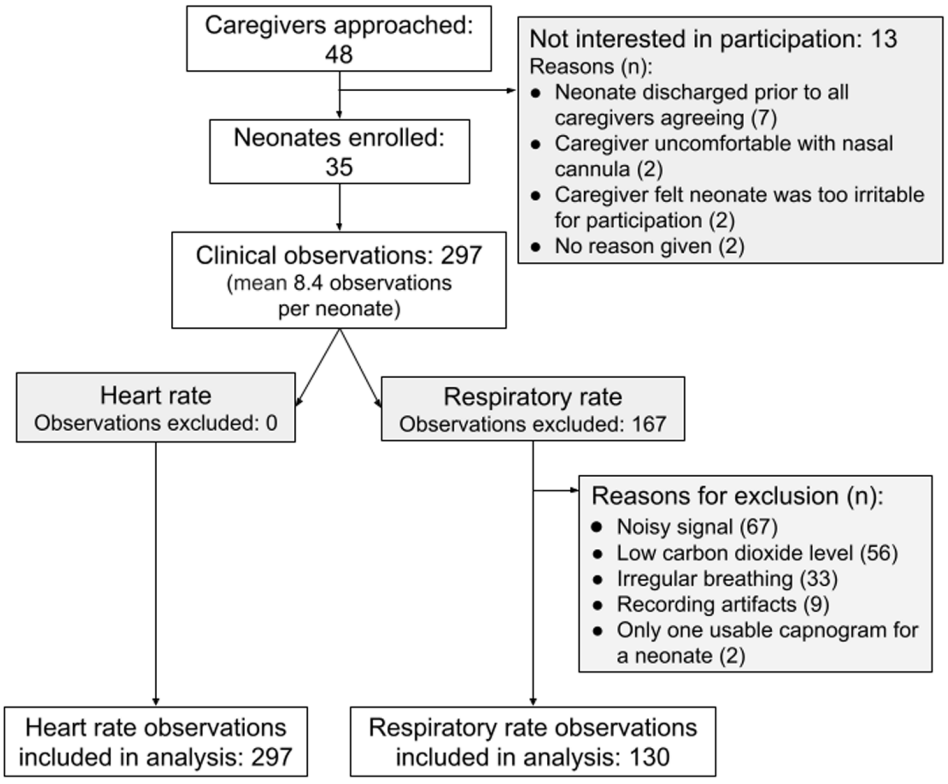


Figure 1. Recruitment flow chart.

Table 4. Bland-Altman analysis of heart rate (HR) and respiratory rate (RR) methods.

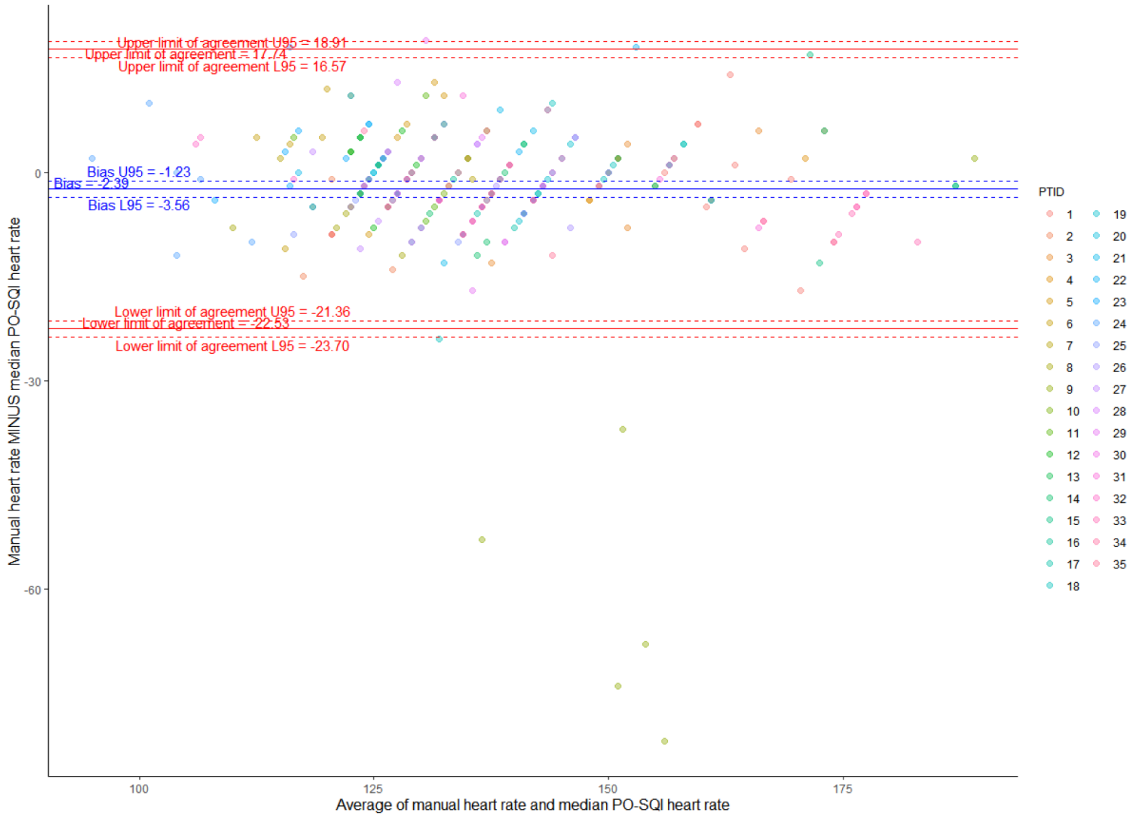
	Bias (normalized)	95% upper/ lower limits of agreement	Spread of 95% limits of agreement (normalized)	Root-mean-square deviation (normalized)
Heart rate				
Manual HR vs median pulse oximetry signal quality index HR	-2.39 (-1.8%)	-22.53/17.74	40.27 (29.6%)	10.5 (7.7%)
Manual HR vs median pulse oximetry signal quality index HR (outlier neonate removed)	-1.4 (-1.0%)	-13.71/10.97	24.67 (18.2%)	6.4 (4.7%)
Respiratory rate				
Manual RR vs algorithm-derived median RR	-3.16 (-6.6%)	-12.1/5.8	17.9 (37.3%)	5.5 (11.4%)
Manual RR vs algorithm-derived RR count	-0.52 (-1.1%)	-2.7/1.66	4.37 (9.1%)	1.2 (2.5%)

The repeatability was moderate with approximately 66% (95 CI 47%-79%) of the manual RR variability due to differences between neonates (Table 5, Figure 3C). Since the 95% CI crossed 50%, the amount of between- and within-neonate variability appeared similar, with neither one resulting in significantly more variability than the other.

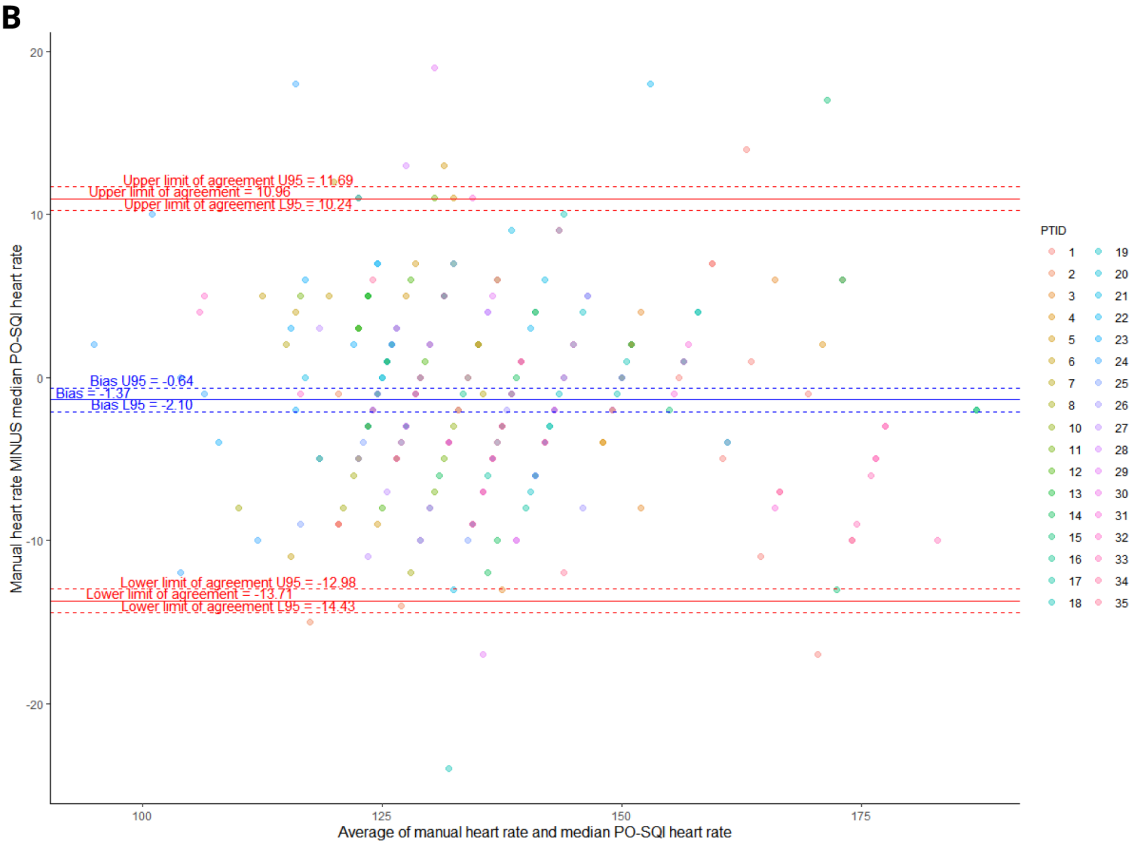
Discussion

This reference technology clinical verification study showed minimal measurement bias with a wide spread of 95% upper and lower LOAs and similar repeatability compared with manual clinical measurements. The agreement results allowed us to identify practical HR and RR thresholds for our subsequent

A


Upper limit of agreement U95 = 18.91
Upper limit of agreement = 17.74
Upper limit of agreement L95 = 16.57
Bias U95 = -1.23
Bias = -2.39
Bias L95 = -3.56
Lower limit of agreement U95 = -21.36
Lower limit of agreement = -22.53
Lower limit of agreement L95 = -23.70
Manual heart rate MINUS median PO-SQI heart rate
Average of manual heart rate and median PO-SQI heart rate
PTID

B


Upper limit of agreement U95 = 11.69
Upper limit of agreement = 10.96
Upper limit of agreement L95 = 10.24
Bias U95 = -0.64
Bias = -1.37
Bias L95 = -2.10
Lower limit of agreement U95 = -12.98
Lower limit of agreement = -13.71
Lower limit of agreement L95 = -14.43
Manual heart rate MINUS median PO-SQI heart rate
Average of manual heart rate and median PO-SQI heart rate
PTID

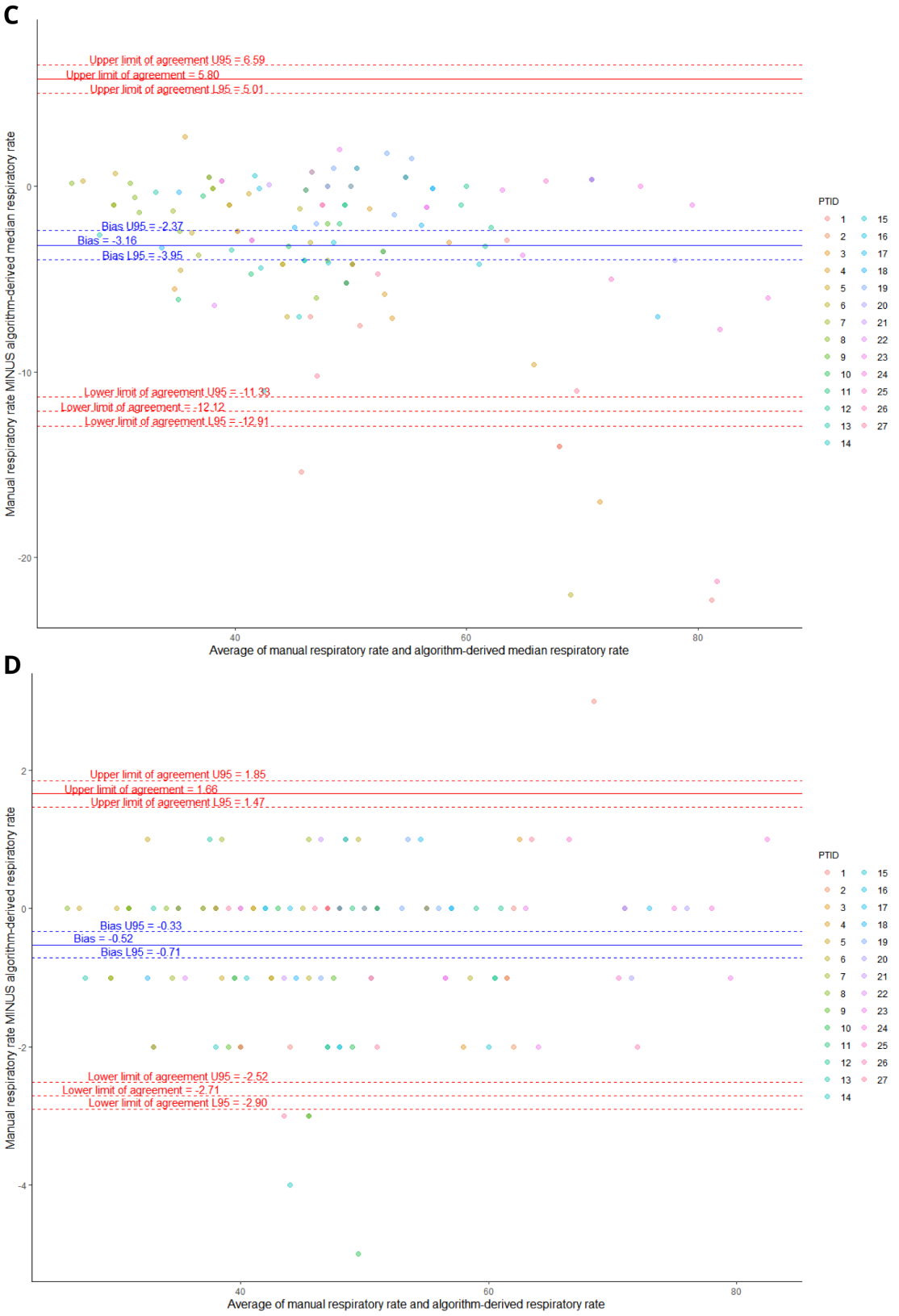


Figure 2. Bland-Altman plots comparing manual heart rate (HR) vs median pulse oximetry signal quality index (PO-SQI) HR for all epochs (**A**), modified manual HR vs median PO-SQI HR with PTID9 removed due to significant outliers (**B**), manual respiratory rate (RR) vs algorithm-derived median RR (**C**), and manual RR vs algorithm-derived RR count (**D**).

Table 5. Repeatability results for heart rate (HR) and respiratory rate (RR) measurements for all included epochs.

	Repeatability[1] (95% Confidence Intervals)
Heart rate (n=297 epochs)	
Manual HR	0.62 (0.47-0.73)
Median pulse oximetry signal quality index HR	0.75 (0.62-0.83)
Respiratory rate (n=130 epochs)	
Manual RR	0.66 (0.47-0.79)
Algorithm-derived median RR	0.50 (0.28-0.67)
Algorithm-derived RR count	0.66 (0.46-0.79)

[1] Repeatability = (between-neonate variance/(between-neonate variance + within-neonate variance))

technology comparison evaluation. Specifically, we identified a 30% spread between the 95% upper and lower LOA. These *a priori*-defined thresholds were based on variability observed ten and sixty minutes apart in the same neonate and considered the natural within-neonate physiologic variability. Variability was found to be more marked in some neonates. In part, the 30% spread between 95% upper and lower LOA was selected based on the idea that thresholds should not be more stringent than the observed physiological variability, and in part, based on results from the different averaging methods (manual RR vs algorithm-derived median RR). Given the large difference in results between the two averaging methods, considerable thought should be given prior to choosing an averaging method. A random selection of real clinical data can provide appropriate guidance for selecting suitable neonatal accuracy thresholds.

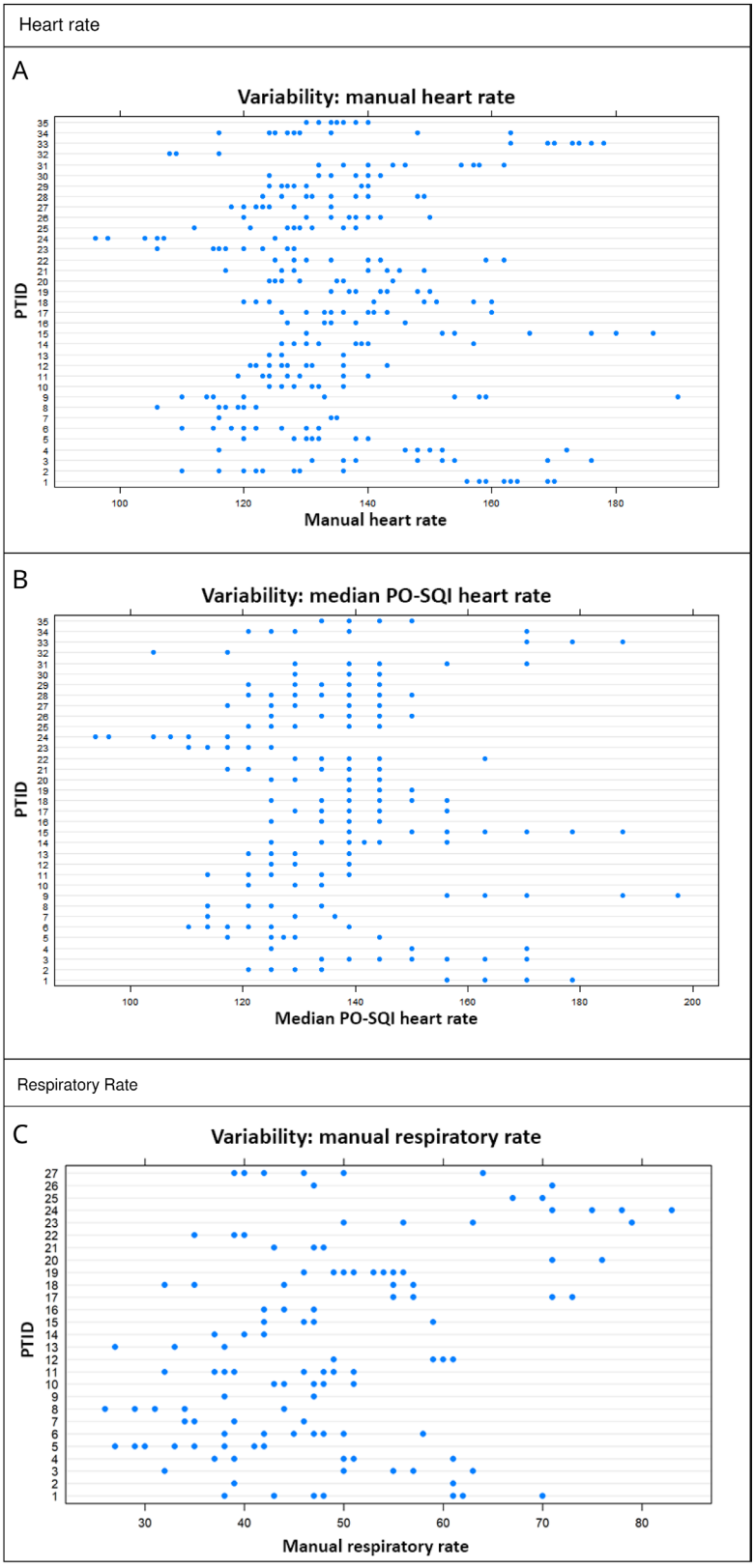
Of note, one neonate (PTID9) significantly impacted the LOA for HR. Five of nine of this neonate's manual HR measurements significantly diverged from the same epoch's PO-SQI HR values and were significantly lower than their mean PO-SQI HR, despite having acceptable signal quality scores. This irregularity suggests a HR reading or data entry error by the study nurse. Removing this neonate's data and re-analyzing it resulted in a smaller bias and tighter LOAs (Figure 2B).

Results from this clinical verification highlight the difficulty with existing performance thresholds. Current United States Food and Drug Administration performance thresholds for HR measurement, based on electrocardiogram measurements, may not be applicable for use in neonates or when using photoplethysmography for estimating HR[19]. The current UNICEF target product profile for RR measurement technology recommends a ±2 breaths per minute threshold, which may be too stringent even for use in adults[20,21]. Using a ±2 breaths per minute recommendation with our RR data would result in a LOA spread threshold of no more than 5%, which is half the LOA spread of our best performing RR comparison. Furthermore, a ±2 breaths per minute or 5% spread in LOA is smaller than random and natural within-neonate physiologic variability (11.5% in this study [unpublished data]) and would result in unrealistically stringent thresholds.

Selecting a performance threshold is challenging. The threshold cannot be too restrictive or inflexible, thereby stifling innovation and preventing new single or multi-parameter continuous monitors from reaching the market. However, too lax a threshold could result in an inaccurate representation of the underlying physiological state. One key limitation is that the true underlying HR or RR is unknown, regardless of the measurement method[6,22]. The primary goal of this reference technology verification study was to establish *a priori* thresholds as the first step of our technology comparison evaluation while at the same time understanding that the true underlying RR and HR cannot be known and also recognizing the marked physiologic variability between and within neonates.

In this study, we did not attempt to define or detect clinically meaningful events; instead, we focused on describing non-random thresholds that fall outside of normal physiological variability. We defined HR and RR thresholds based on the difference between the 95% upper and lower LOA. Additional studies will be required to determine if these thresholds translate into improved clinical outcomes.

Performance thresholds identified using this method are influenced by the characteristics of the neonates studied, the data selection methods, and the number of comparisons. For this reason, the thresholds we identified may not be applicable in different neonate cohorts, such as those receiving mechanical ventilation or immediately following birth, among others. Variability will be influenced by disturbances in the environment such as routine procedures, feeding, noise, and time of day. To minimize variability in our data set, we used only RR epochs that appeared to be regular based on visual inspection. Although these segments were selected based on predefined


Heart rate
A
Variability: manual heart rate
PTID
Manual heart rate
B
Variability: median PO-SQI heart rate
PTID
Median PO-SQI heart rate
Respiratory Rate
C
Variability: manual respiratory rate
PTID
Manual respiratory rate

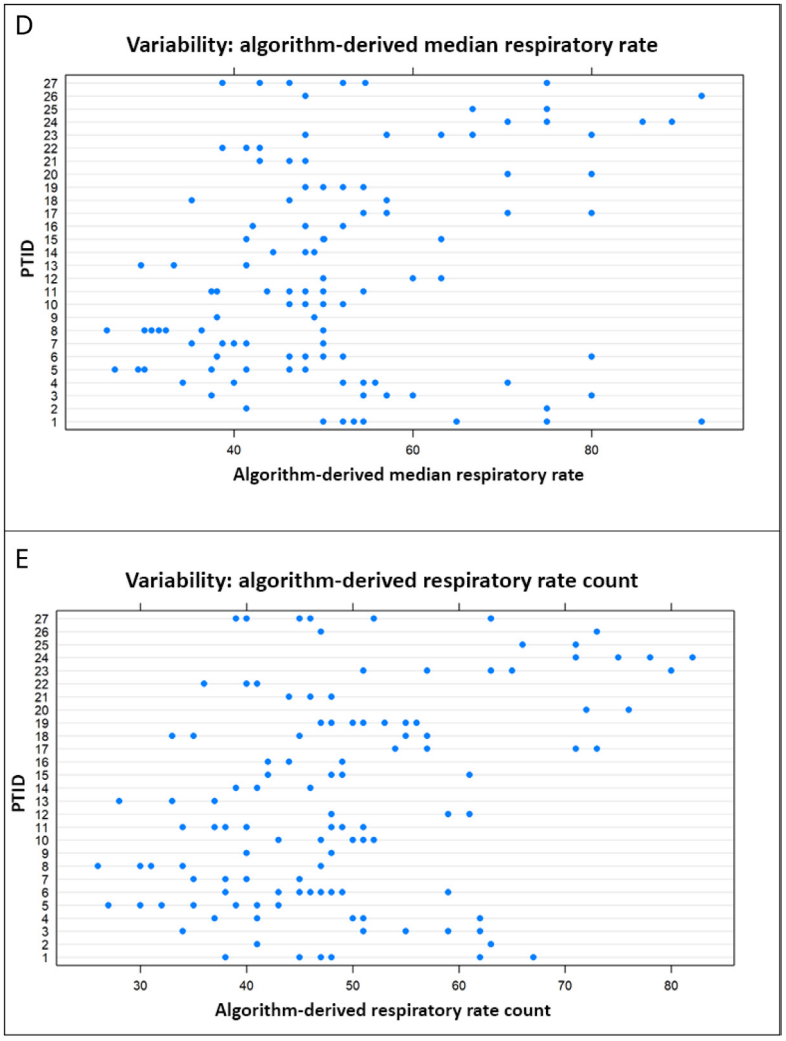


Figure 3. Variability plots (vertical for between-neonate variability, horizontal for within-neonate variability). Manual heart rate (HR) between-neonate variability accounts for 62% of total variability (**A**); median pulse oximetry signal quality index (PO-SQI) HR between-neonate variability accounts for 75% of total variability (**B**); manual respiratory rate (RR) between-neonate variability accounts for 66% of total variability (**C**); algorithm-derived median RR between-neonate variability accounts for 50% of total variability (**D**); and algorithm-derived RR count between-neonate variability accounts for 66% of total variability (**E**).

criteria, a majority (167/297) were discarded as the extreme variability seen in some recordings would have made reproducible manual counting of breaths impossible. We have previously demonstrated acceptable agreement between ECG derived HRV and PPG derived HRV in children with an appropriate sampling rate of the PPG. This should be validated in neonates using an ECG[23].

Conclusion

Appropriate clinical thresholds should be selected *a priori* when performing accuracy comparisons for HR and RR. The magnitude and importance of sample size, as well as within-neonate variability requires further investigation. A larger sample size could allow the development of an error model that more clearly describes the error due to various factors such as the measurement technology, averaging method, the observer, and the natural variability of neonatal HR and RR. We strongly support the creation of international standards for technology comparison studies in neonates. These standards should include thresholds for HR and RR based on the specific neonatal population studied and provide details of the experimental conditions, data selection methods, and analysis methods used. Together, such standards would lay the groundwork for a robust continuous monitor comparison field.

Data availability

Underlying data

Dryad: Identification of thresholds for accuracy comparisons of heart rate and respiratory rate in neonates. https://doi.org/10.5061/dryad.1c59zw3vb[18].

This project contains the following underlying data:

- Coleman-2021-ETNA-DemographicData.csv
- Raw data folder (contains all raw capnography and pleth data)
- Coleman-2021-ETNA-ProcessedPulseValues.csv
- Coleman-2021-ETNA-ProcessedRespirationValues.csv

Acknowledgments

We thank Dorothy Chomba, Millicent Parsimei, Annah Kasyoka, and the dedicated staff at Aga Khan University-Nairobi Hospital for providing patient care, and the neonatal participants, their caregivers, and the local community in Nairobi, Kenya, for their participation.

References

1. United Nations Inter-agency Group for Child Mortality Estimation (UN IGME): **Levels & Trends in Child Mortality: Report 2020, Estimates developed by the UN Inter-agency Group for Child Mortality Estimation.** New York: United Nations Children's Fund, 2020.
Reference Source
2. Fairchild KD, Schelonka RL, Kaufman DA, *et al.*: **Septicemia mortality reduction in neonates in a heart rate characteristics monitoring trial.** *Pediatr Res.* 2013; **74**(5): 570–5.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Warburton A, Monga R, Sampath V, *et al.*: **Continuous pulse oximetry and respiratory rate trends predict short-term respiratory and growth outcomes in premature infants.** *Pediatr Res.* 2019; **85**(4): 494–501.
PubMed Abstract | Publisher Full Text
4. Kumar N, Akangire G, Sullivan B, *et al.*: **Continuous vital sign analysis for predicting and preventing neonatal diseases in the twenty-first century: big data to the forefront.** *Pediatr Res.* 2020; **87**(2): 210–20.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Harris BU, Char DS, Feinstein JA, *et al.*: **Accuracy of Pulse Oximeters Intended for Hypoxemic Pediatric Patients.** *Pediatr Crit Care Med.* 2016; **17**(4): 315–20.
PubMed Abstract | Publisher Full Text
6. Ansermino JM, Dumont G, Ginsburg AS: **How Uncertain Is Our Reference Standard for Respiratory Rate Measurement?** *Am J Respir Crit Care Med.* 2019; **199**(8): 1036–7.
PubMed Abstract | Publisher Full Text
7. Ginsburg AS, Nkwopara E, Macharia W, *et al.*: **Evaluation of non-invasive continuous physiological monitoring devices for neonates in Nairobi, Kenya: a research protocol.** *BMJ Open.* 2020; **10**(4): e035184.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Goldsack JC, Coravos A, Bakker JP, *et al.*: **Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs).** *NPJ Digit Med.* 2020; **3**: 55.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Lee HJ, Choi JH, Min SJ, *et al.*: **Comparison of the Clinical Performance between Two Pulse Oximeters in NICU: Nellcor N-595(R) versus Masimo SET(R).** *Journal of the Korean Society of Neonatology.* 2010; **17**(2): 245–249.
Publisher Full Text
10. Singh JKSB, Kamlin COF, Morley CJ, *et al.*: **Accuracy of pulse oximetry in assessing heart rate of infants in the neonatal intensive care unit.** *J Paediatr Child Health.* 2008; **44**(5): 273–5.
PubMed Abstract | Publisher Full Text
11. Hay WW Jr, Rodden DJ, Collins SM, *et al.*: **Reliability of conventional and new pulse oximetry in neonatal patients.** *J Perinatol.* 2002; **22**(5): 360–6.
PubMed Abstract | Publisher Full Text
12. World Health Organization: **Integrated management of childhood illness: caring for newborns and children in the community**. 2011.
Reference Source
13. Harris PA, Taylor R, Thielke R, *et al.*: **Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support.** *J Biomed Inform.* 2009; **42**(2): 377–81.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Karlen W, Ansermino JM, Dumont G: **Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications.** *Annu Int Conf IEEE Eng Med Biol Soc.* 2012; **2012**: 3131–4.
PubMed Abstract | Publisher Full Text
15. R Core Team: **R: A language and environment for statistical computing**.
Reference Source
16. Stoffel MA, Nakagawa S, Schielzeth H: **rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models.** *Methods Ecol Evol.* 2017; **8**(11): 1639–44.
Publisher Full Text
17. Bland JM, Altman DG: **Measuring agreement in method comparison studies.** *Stat Methods Med Res.* 1999; **8**(2): 135–60.
PubMed Abstract | Publisher Full Text
18. Coleman J: **Identification of thresholds for accuracy comparisons of heart rate and respiratory rate in neonates.** Dryad, Dataset, 2021.
http://www.doi.org/10.5061/dryad.1c59zw3vb
19. American National Standards Institute, Inc: **Cardiac monitors, heart rate meters, and alarms.** {Association for the Advancement of Medical Instrumentation} 2002.
20. UNICEF: **Target Product Profile - Respiratory Rate Monitor / Apnea Monitor**. 2020.
Reference Source
21. Ermer S, Brewer L, Orr J, *et al.*: **Comparison of 7 Different Sensors for Detecting Low Respiratory Rates Using a Single Breath Detection Algorithm in Nonintubated, Sedated Volunteers.** *Anesth Analg.* 2019; **129**(2): 399–408.
PubMed Abstract | Publisher Full Text
22. Ginsburg AS, Lenahan JL, Izadnegahdar R, *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): 1116–27.
PubMed Abstract | Publisher Full Text
23. Dehkordi P, Garde A, Karlen W, *et al.*: **Pulse rate variability compared with Heart Rate Variability in children with and without sleep disordered breathing.** *Annu Int Conf IEEE Eng Med Biol Soc.* 2013; **2013**: 6563–6566.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ✗ ✓

Version 2

Reviewer Report 29 November 2021

https://doi.org/10.21956/gatesopenres.14653.r31258

✓ **AbdelKebir Sabil**
Cloud Sleep Lab, Paris, France

The method and the results of this study are of interest and the authors seem to have answered the concerns expressed by the reviewers after the initial submission. Despite the fact that I did not participate in the first review of the manuscript, I believe that it would have been interesting if data comparing heart rate variability measured using ECG to pulse rate variability measured with the oximeter. In that case, indices in the time domain (SDNN and RMSSD) and frequency domain (LF, HF and LF/HF) should be compared using different methods. The other comment I have is that the sample size is a bit too small.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Technology clinical validation expert.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 09 November 2021

https://doi.org/10.21956/gatesopenres.14653.r31254

Robert E. Kearney
Department of Biomedical Engineering, McGill University, Montreal, QC, Canada

In this work the authors compare manual measures of heart rate and respiratory with measures obtained through the automated analysis of pulse oximetry and capnograph signals. The stated objective is to determine the threshold to be used in evaluating the accuracy of new continuous monitors. The major finding was that the different methods had low bias but large random errors. Given this finding it is not clear how the results of this paper contribute to its stated objectives. Indeed, the authors conclude that appropriate clinical thresholds should be selected a priori.
I have several concerns with the paper as it stands mostly related to the methods:

1. I had difficulty understanding exactly what measures were being compared. As I understand it: For heart rate the measures were: (1) The average heart rate over a 60 second period computed as: The reciprocal of the number of heart beats manually over the period. (2) The median of the reciprocal of the beat to beat to beat interval computed by the Masimo device. For respiratory rate the two measures were (1) The reciprocal of the number of breaths in period determined by manual analysis of the capnograph and (2) The reciprocal of the median breath duration computed over the same period using a computer algorithm. This is correct, then the authors are not actually comparing measures of heart rate and respiratory rate but rather measures of their average values computed over a one-minute period. A difficulty with this is that this provides no measure of the accuracy of measures of either HR or RR variability. The authors need to make it clear exactly what they were comparing, justify their choice, and explain why they used a one-minute period.

2. The authors use confidence intervals and limits of agreement derived from the Bland Altman plots to assess the differences between measures. My understanding is that the validity of these measures depends on the assumption that the differences between measures are normally distributed. Did the authors validate this assumption?

3. The automatic analysis of the capnograph was done using an algorithm developed by the authors and described in a conference paper. There is no discussion of how this algorithm was validated or what its expected accuracy was.

4. The paper would be greatly improved by the inclusion of a figure showing a typical data record with manual markings and a clear definition of what the various measures were.

5. One of the authors' conclusions is that a wider spread between the LOA values should be allowed to account for intra- and inter-neonate variability. It is not clear to me why this should be.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biomedical signal and system analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 01 November 2021

https://doi.org/10.21956/gatesopenres.14653.r31256

Kevin Baker
[1] Malaria Consortium, London, UK
[2] Karolinska Institute, Stockholm, Sweden

Thanks for the opportunity to review this important work. It is a very well written piece and congratulations to the authors on describing this work so clearly.

I think this work builds on a number of previous studies and as there are so many groups working on RR at the moment I think it would have be useful to situate this work in relation to the previous studies and articles (listed below) and discussions that have taken place, I think by showing that these results are similar it strengthens the arguments around what the apriori thresholds should be. Again these were discussed at a UNICEF meeting in 2019 and this could be referenced also - to show that these findings match global discussions.

In measuring RR we know that movement has a huge impact on the variability of RR - this is not well described in the piece. The authors mention "To minimize variability in our data set, we used only RR epochs that appeared to be regular based on visual inspection. Although these segments were selected based on predefined criteria, a majority (167/297) were discarded as the extreme variability seen in some recordings would have made reproducible manual counting of breaths impossible". Does this mean you removed RR epochs or instances where there was a lot of movement? While I agree that it is good to reduce variability I would be concerned that by removing the highly variable epochs the authors are not reflecting a true RR. Apologies if I misunderstood the methods here.

Articles to reference in the background:

- Carina *et al.* (2021[1]).
- Stratil *et al.* (2021[2]).
- Baker *et al.* (2019[3]).

References

1. Carina K, Kevin B, Rebecca N, Bassat Q, et al.: Back to Basics in Paediatric Pneumonia—Defining a Breath and Setting Reference Standards to Innovate Respiratory Rate Counting. *Journal of Tropical Pediatrics*. 2021; **67** (1). Publisher Full Text
2. Stratil A, Ward C, Habte T, Maurel A, et al.: Evaluating the Interrater Agreement and Acceptability of a New Reference Tool for Assessing Respiratory Rate in Children under Five with Cough and/or Difficulty Breathing. *Journal of Tropical Pediatrics*. 2021; **67** (2). Publisher Full Text
3. Baker K, Alfvén T, Mucunguzi A, Wharton-Smith A, 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**: 20-30 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Pneumonia malaria child health

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 18 October 2021

https://doi.org/10.21956/gatesopenres.14653.r31230

?

Gordon B. Drummond
Department of Anaesthesia, Critical Care, and Pain Medicine, University of Edinburgh, Edinburgh, UK

I've now developed one further question about the data measurement process. It is this:

If respiratory rate was measured from exactly the same epoch, and standard criteria were used for defining a breath, is it necessarily also true that exactly the same breaths were counted? I suggest this is not the case, and this is for the reason the authors imply, but do not elucidate: the monitor-derived value is calculated from "past values" and displayed in the "present". The observer based measure is then taken from events that happen from that time forward. I would suggest a simple test of the monitors, using a simulated sample of waveforms, whose duration could be abruptly changed, say from 2 seconds to 2.5 seconds, would substantially elucidate the capacity of the device to reflect the "real now" signal that the observer has been set to observe. A diagram to show the relative times of what is measured by a monitor, and an observer, relative to the "reference mark" time, would be helpful for the reader to grasp these concepts.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Nov 2021

Jesse Coleman, Aga Khan University Hospital, Nairobi, Kenya

Dear Dr. Drummond,

Thank you for providing the opportunity to respond to your query on our updated version of the manuscript titled “Identification of thresholds for accuracy comparisons of heart rate and respiratory rate in neonates” at *Gates Open Access*. We appreciate your ongoing effort to review the manuscript and are grateful for your further comment. Below is our response to your query:

Query: *If respiratory rate was measured from exactly the same epoch, and standard criteria were used for defining a breath, is it necessarily also true that exactly the same breaths were counted? I suggest this is not the case, and this is for the reason the authors imply, but do not elucidate: the monitor-derived value is calculated from "past values" and displayed in the "present". The observer based measure is then taken from events that happen from that time forward. I would suggest a simple test of the monitors, using a simulated sample of waveforms, whose duration could be abruptly changed, say from 2 seconds to 2.5 seconds, would substantially elucidate the capacity of the device to reflect the "real now" signal that the observer has been set to observe. A diagram to show the relative times of what is measured by a monitor, and an observer, relative to the "reference mark" time, would be helpful for the reader to grasp these concepts.*

Response: Thank you for highlighting this tricky aspect of breath identification and respiratory rate comparison. We were able to match the exact breath. Our team had access to the raw (instantaneous) CO_2 waveform data, recorded at approximately 20 Hz. We only analyzed the raw waveform data and counted the number of breaths. Furthermore, each 60-second epoch was isolated; no data from before or after the epoch was included in any calculation or analysis. Rather, individual breaths were counted using two different breath

counting methods; 1, Study team members manual breath counting from capnograms using standardized breath identification rules, and 2, Algorithm-derived breath identification developed in MATLAB. Your suggestion about simulated respiratory rate comparisons may be an alternative if we did not have this very precise breath identification or if some method of filtering or averaging of RR was used.

Competing Interests: None

Version 1

Reviewer Report 03 August 2021

https://doi.org/10.21956/gatesopenres.14469.r30945

Gordon B. Drummond
Department of Anaesthesia, Critical Care, and Pain Medicine, University of Edinburgh, Edinburgh, UK

The exact hypothesis of this study is hard to discern. In their abstract and introduction, the authors imply that innovative, *non-invasive measurement* technologies that use advanced measures of vital signs such as heart rate variation and transient deceleration (citation 2) can be used to improve outcome in infants in resource-constrained settings such as low and middle income countries, but the paper then describes a comparison of nurse observation with continuous measures available from electronic monitors, with the stated aim of defining the accuracy of methods to continuously measure physiological events. Such comparisons have been done, and they cite a substantial review (citation 4).

The introduction then ends with this statement of the study aim: “the clinical reference technology verification processes conducted to determine appropriate heart rate (HR) and respiratory rate (RR) thresholds in subsequent accuracy comparisons.”
However the methods then state the aim is “to identify the natural variation in neonatal HR and RR in order to identify appropriate accuracy thresholds for use in an accuracy comparison of MCPM technologies.”
So, we have at least three alternative study aims: the third I’d consider to be the most useful aim, comparing MCPM methods: unlikely to be answered when comparing clinicians with monitors, but could be answered with the data gathered.

At this point, I felt that some sensible and more exact definitions are required, for words such as accuracy, repeatability, agreement, threshold, precision perhaps – as stated in citation 6, by two of the authors of the present paper.

What is “Repeatability”? If we accept that the result of a 60 second counting period will differ, from one observation to the next, because the components of the measure (the duration of each breath, or the interval between photoplethysmograph pulse waves) are randomly different, then the only mechanism available to improve the *estimate* of the overall frequency is to increase the size of the sample: this is the law of large numbers, a statistical rule that has been known for several centuries in one form or another.

Bland and Altman, when first introducing their extremely popular method, used an example of spirometry: a single measure made first with one device and then with an alternative device. It’s quite possible that two repeat FVC manoeuvres with the same device would differ: within subject variation. This is a more substantial problem in this study, as the authors state:
“Furthermore, a ±2 breaths per minute or 5% spread in LOA is smaller than random and natural within-neonate physiologic variability (11.5% in this study [unpublished data]) and would result in unrealistically stringent thresholds”. The degree of within subject variation is evident also from Figure 3. The phenomenon was noted by Simoes *et al.* (1991[1]).

So we have a small number of intrinsically variable events. So, for a fair comparison of two methods, a necessary requirement is to ensure that the events being measured are the same, exactly the same sample has been taken. If the pulse-wave derived rate from the machine is of a different series of waves (i.e the time period is not EXACTLY the same) than those counted by the nurse, they are already going to be affected by within subject variation as well as the variation between the methods. The methods state: “Manual measures were every 10 minutes for the first hour and once per hour of participation thereafter: were the manual and monitor measures exactly timed to coincide? And, was there any time trend in the patients studied for longer times?

Of course, Bland and Altman had to subsequently refine their method, to separately account for repeated measures in multiple subjects, and at the same time they introduced the concept of confidence intervals for the limits of agreement. Looking at figure 3, there’s a lot of variation: it would be helpful to plot the CI for the LOA on the Bland and Altman plots.
However, I would suggest that the most useful thing to do would be to carefully analyse repeated random samples from the electronic records, looking at precise time intervals, so that the intrinsic variation could be quantified, and study how different sample sizes might affect reliability of the rate values. We have done this for respiratory rate in acutely ill adults (Drummond *et al.*, 2020[2]). Using 30 second periods of observation gave an interquartile range of respiratory rates of 3.4 breath/minute, whereas samples taken for 120 seconds had an IQR of 2.5. Using the techniques the authors describe here, why not sample for 5 minutes?

Availability of these records would be very useful to other workers! More analysis of the monitor records is also important since it appears that rate is not, in itself, perhaps the most important signal. For example, others have found that short-term heart and respiratory rate variability make a significant contribution to illness scoring systems (Saria *et al.*, 2010[3]).

Small points:

- Abstract: “Automated measurement technologies typically use median values”. I’m afraid that my experience is that manufacturers of monitors rarely tell what they use: some sort of exponential averaging or filtering seems more likely. It would be good to have this statement substantiated (if possible).

- Abbreviations make reading difficult. “multiparameter continuous physiological monitoring (MCPM) technologies” is subsequently used as MCPM technologies (17 characters) throughout the paper. Why not just use “continuous monitors” (19 characters)?

- Pulse plethysmography may not be an accurate measure of heart rate variability. ECG monitoring might be better. I realise that ECG has its drawbacks!

References

1. Simoes EA, Roark R, Berman S, Esler LL, et al.: Respiratory rate: measurement of variability over time and accuracy at different counting periods.*Arch Dis Child*. 1991; **66** (10): 1199-203 PubMed Abstract | Publisher Full Text
2. Drummond GB, Fischer D, Arvind DK: Current clinical methods of measurement of respiratory rate give imprecise values.*ERJ Open Res*. 2020; **6** (3). PubMed Abstract | Publisher Full Text
3. Saria S, Rajani AK, Gould J, Koller D, et al.: Integration of early physiological responses predicts later illness severity in preterm infants.*Sci Transl Med*. 2010; **2** (48): 48ra65 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Expertise in respiratory monitoring and data analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 27 Sep 2021

Jesse Coleman, Aga Khan University Hospital, Nairobi, Kenya

We thank Dr. Drummond for their time and effort in providing valuable feedback and we are grateful to you for the insightful comments. We have been able to incorporate changes to reflect most of the suggestions you provided. Our responses to the points raised are below:

Point 1: *The exact hypothesis of this study is hard to discern. In their abstract and introduction, the authors imply that innovative, non-invasive measurement technologies that use advanced measures of vital signs such as heart rate variation and transient deceleration (citation 2) can be used to improve outcome in infants in resource-constrained settings such as low and middle income countries, but the paper then describes a comparison of nurse observation with continuous measures available from electronic monitors, with the stated aim of defining the accuracy of methods to continuously measure physiological events. Such comparisons have been done, and they cite a substantial review (citation 4).*
The introduction then ends with this statement of the study aim: "the clinical reference technology verification processes conducted to determine appropriate heart rate (HR) and respiratory rate (RR) thresholds in subsequent accuracy comparisons."
However the methods then state the aim is "to identify the natural variation in neonatal HR and RR in order to identify appropriate accuracy thresholds for use in an accuracy comparison of MCPM technologies."
So, we have at least three alternative study aims: the third I'd consider to be the most useful aim, comparing MCPM methods: unlikely to be answered when comparing clinicians with monitors, but could be answered with the data gathered.
Response: Thank you for identifying that the language used to introduce the topic might not align with the stated goals of the manuscript. The reviewer is correct that the 3rd aim is what we have addressed. We will be modifying the language in the introduction section to read "**This study describes the verification processes we conducted with a clinical reference technology in order to determine appropriate heart rate (HR) and respiratory rate (RR) accuracy thresholds to use in subsequent new patient monitoring technology accuracy comparisons.**"

Point 2: *At this point, I felt that some sensible and more exact definitions are required, for words such as accuracy, repeatability, agreement, threshold, precision perhaps – as stated in citation 6, by two of the authors of the present paper*.
Response: We agree with your recommendation to clarify the definitions of some keywords. We will be adding definitions for the suggested terms to the definitions table to read as follows:
Accuracy: The closeness a measured value is from the true value
Repeatability: The closeness of the results of successive measurements of the same measure
Agreement (between measures): The consistency between two sets of measurements
Accuracy Threshold: A pre-specified value used to determine if a set of measurements has achieved a sufficient accuracy when compared with a reference value
Precision: The closeness of measurements to each other

Point 3: *What is "Repeatability"? If we accept that the result of a 60 second counting period will*

differ, from one observation to the next, because the components of the measure (the duration of each breath, or the interval between photoplethysmograph pulse waves) are randomly different, then the only mechanism available to improve the estimate of the overall frequency is to increase the size of the sample: this is the law of large numbers, a statistical rule that has been known for several centuries in one form or another.
Bland and Altman, when first introducing their extremely popular method, used an example of spirometry: a single measure made first with one device and then with an alternative device. It's quite possible that two repeat FVC manoeuvres with the same device would differ: within subject variation. This is a more substantial problem in this study, as the authors state: "Furthermore, a ±2 breaths per minute or 5% spread in LOA is smaller than random and natural within-neonate physiologic variability (11.5% in this study [unpublished data]) and would result in unrealistically stringent thresholds". The degree of within subject variation is evident also from Figure 3. The phenomenon was noted by Simoes et al. (19911).
So we have a small number of intrinsically variable events. So, for a fair comparison of two methods, a necessary requirement is to ensure that the events being measured are the same, exactly the same sample has been taken. If the pulse-wave derived rate from the machine is of a different series of waves (i.e the time period is not EXACTLY the same) than those counted by the nurse, they are already going to be affected by within subject variation as well as the variation between the methods. The methods state: "Manual measures were every 10 minutes for the first hour and once per hour of participation thereafter: were the manual and monitor measures exactly timed to coincide?
Response: You have raised a critically important issue of synchronization. We apologize that we did not emphasize the importance in the original draft. We have updated our methods section to clearly describe the synchronization methods we used to ensure that all data was precisely temporally aligned. The new wording will read as follows: "**To ensure temporal alignment between measurements, HR and RR epochs were synchronized across source data devices. For HR, alignment was done using a timestamp in REDCap that was set by the study nurse as HR counting was initiated. Before analysis, this timestamp was synchronized with the same timestamp in the custom Android application. Both the REDCap and Android servers were connected via the internet to a Network Time Protocol (NTP) server. Alignment of RR epochs was based on the Android application timestamp. All RR waveforms were compared visually to further ensure epoch synchronization.**" With the definition and clarification made, we feel that testing the repeatability and agreement of the two methods is reasonable.

Point 4: *And, was there any time trend in the patients studied for longer times?*
Response: Thank you for asking this question. We do have respiratory rate data on patients studied for longer times that has been submitted and is currently under review elsewhere.

Point 5: *Of course, Bland and Altman had to subsequently refine their method, to separately account for repeated measures in multiple subjects, and at the same time they introduced the concept of confidence intervals for the limits of agreement. Looking at figure 3, there's a lot of variation: it would be helpful to plot the CI for the LOA on the Bland and Altman plots.*
Response: Thank you for the helpful suggestion. We will be updating the Bland and Altman plots to include the CI for the LOA throughout.

Point 6: *However, I would suggest that the most useful thing to do would be to carefully analyse*

repeated random samples from the electronic records, looking at precise time intervals, so that the intrinsic variation could be quantified, and study how different sample sizes might affect reliability of the rate values. We have done this for respiratory rate in acutely ill adults (Drummond et al., 2020).
Using 30 second periods of observation gave an interquartile range of respiratory rates of 3.4 breath/minute, whereas samples taken for 120 seconds had an IQR of 2.5. Using the techniques the authors describe here, why not sample for 5 minutes?
Response: Thank you for this suggestion. It would have been important to look at repeated random samples if we did not have temporally aligned data. However, in the case of our study, it is unnecessary because we can compare measurements from the same epoch. We do have a manuscript in preparation that will address this clinical issue of counting. Contrary to your suggestion we found the agreement deteriorated in observations over one minute. In this manuscript, we focus on the thresholds we would consider appropriate for a subsequent method comparison study.

Point 7: *Availability of these records would be very useful to other workers ! More analysis of the monitor records is also important since it appears that rate is not, in itself, perhaps the most important signal. For example, others have found that short-term heart and respiratory rate variability make a significant contribution to illness scoring systems (Saria et al., 2010).*
Response: Thank you for this suggestion. We completely agree and one of the reasons for the current research is to facilitate the introduction of high-quality vital sign monitors in resource-constrained settings. Our team works closely with manufacturers, who we will recommend include notifications of short-term HR and RR variability. Also, All the data will be made available for other researchers to allow them to address these important considerations.

Small points
Small point 1: *Abstract: "Automated measurement technologies typically use median values". I'm afraid that my experience is that manufacturers of monitors rarely tell what they use: some sort of exponential averaging or filtering seems more likely. It would be good to have this statement substantiated (if possible).*
Response: You are correct. A median is a simple filter and other filters may also be used and would be considered a trade secret. I think the key point here is that a count of breaths is likely to be very different from any filter method that might be used. We will be adjusting the text to read "**typically use a smoothing or averaging filter**"

Small point 2: *Abbreviations make reading difficult. "multiparameter continuous physiological monitoring (MCPM) technologies" is subsequently used as MCPM technologies (17 characters) throughout the paper. Why not just use "continuous monitors" (19 characters)?*
Response: Thank you for this suggestion. We understand that acronyms can be difficult to read. To facilitate readability we will integrate your suggested terminology to read " **continuous monitors**" throughout.

Small point 3: *Pulse plethysmography may not be an accurate measure of heart rate variability. ECG monitoring might be better. I realise that ECG has its drawbacks!*
Response: We would agree that there is a difference between heart variability derived from ECG and that derived from pulse plethysmography. We have previously compared the

measures of heart rate variability between ECG and plethysmography in children. This agreement is dependent on an appropriate sampling rate of the plethysmogram. Noting your comment, we will add the following text to the discussion: "**Pulse plethysmography may not be an accurate measure of HR variability due to innate technology limitations. Future studies looking at HR variability should consider using ECG monitoring, despite having its own limitations.[1]**"

1. Reference: Dehkordi, P., Garde, A., Karlen, W., Wensley, D., Ansermino, J. M., & Dumont, G. A. (2013). Pulse rate variability compared with Heart Rate Variability in children with and without sleep disordered breathing. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, 6563–6566. https://doi.org/10.1109/EMBC.2013.6611059

Competing Interests: None