BMJ Paediatrics Open

Multi-center paired non-inferiority study of the cardiorespiratory monitoring performance of the wireless and non-adhesive Bambi® belt measuring diaphragm activity in neonates: Study Protocol

Journal:	*BMJ Paediatrics Open*
Manuscript ID	bmjpo-2022-001430
Article Type:	Protocol
Date Submitted by the Author:	27-Jan-2022
Complete List of Authors:	Scholten, Anouk; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute Zhuozhao, Zhan; Eindhoven University of Technology, Department of Mathematics and Computer Science Niemarkt, Hendrik; Máxima Medisch Centrum, Neonatology Vervoorn, Marieke; Máxima Medisch Centrum, Neonatology van Leuteren, Ruud; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute de Jongh, Frans; Amsterdam UMC Locatie AMC, Neonatology; University of Twente, Faculty of Science and Technology van Kaam, Anton; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie VUmc, Neonatology van den Heuvel, Edwin; Eindhoven University of Technology, Department of Mathematics and Computer Science Hutten, G. Jeroen; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie VUmc, Neonatology
Keywords:	Neonatology, Technology

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

MULTI-CENTER PAIRED NON-INFERIORITY STUDY OF THE CARDIORESPIRATORY MONITORING PERFORMANCE OF THE WIRELESS AND NON-ADHESIVE BAMBI® BELT MEASURING DIAPHRAGM ACTIVITY IN NEONATES: STUDY PROTOCOL

Anouk W.J. Scholten[1,2*†], Zhan Zhuozhao[3†], Hendrik J. Niemarkt[4], Marieke Vervoorn[4], Ruud W. van Leuteren[1,2], Frans H. de Jongh[1,5], Anton H. van Kaam[1,6], Edwin R. van den Heuvel[3], G. Jeroen Hutten[1,6]

[1]*Department of neonatology, Emma Children's Hospital, AmsterdamUMC, University of Amsterdam, Amsterdam, The Netherlands*

[2]*Amsterdam Reproduction & Development research institute, Amsterdam, The Netherlands*

3Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[4]*Department of neonatology, Màxima Medical Center, Veldhoven, The Netherlands*

[5]*Faculty of Science and Technology, University of Twente, Enschede, The Netherlands*

[6]*Department of neonatology, Emma Children's Hospital, AmsterdamUMC, Vrije Universiteit, Amsterdam, the Netherlands*

† These authors have contributed equally to this work and share first authorship

CORRESPONDING AUTHOR:

Anouk W.J. Scholten

Department of Neonatology

Emma Children's Hospital, AmsterdamUMC, location AMC

Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

Phone: +31 20 5663675

Email: a.w.j.scholten@amsterdamumc.nl

ABSTRACT

Introduction: Cardiorespiratory monitoring is used in the Neonatal Intensive Care Unit (NICU) to assess the clinical status of newborn infants and detect critical deteriorations in cardiorespiratory function. Currently, heart rate is monitored by electrocardiography (ECG) and respiration by chest impedance (CI). Disadvantages of current monitoring techniques are usage of wired adhesive electrodes which may damage the skin and hinder care. The Bambi® belt is a wireless and non-adhesive alternative that enables cardiorespiratory monitoring by measuring electrical activity of the diaphragm via transcutaneous electromyography (dEMG). A previous study showed feasibility of the Bambi belt and this study compares the belt performance to ECG and CI.

Methods and analysis: This multi-center non-inferiority paired study will be performed in the NICU of the Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital AmsterdamUMC in Amsterdam, the Netherlands. 39 infants in different postmenstrual age groups (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks) will be recruited. These infants will be monitored with the Bambi® belt in addition to standard ECG and CI for 24 h. The primary outcome is the heart rate (HR), studied with three criteria: 1) the agreement in HR measurements between the belt and standard ECG, 2) the detection of cardiac events consisting of bradycardia and tachycardia and 3) the quality of HR-monitoring. The secondary outcome is the RR, studied with the criteria 1) agreement in respiratory rate (RR) trend monitoring, 2) apnea and tachypnea detection, and 3) reliable registrations.

Ethics and dissemination: This protocol was approved by the Medical Ethical Committee of the Màxima Medical Center and the Central Committee for Human Research (CCMO). The MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The results will be presented at conferences and published in peer-reviewed journals.

Trial registration number: NL9480 (www.trialregister.nl)

INTRODUCTION

In the Neonatal Intensive Care Unit (NICU), cardiorespiratory monitoring is crucial to assess clinical condition and to timely detect and treat frequently occurring cardio-respiratory events to prevent morbidity and mortality.(1, 2) To date, this is performed by measuring the electrocardiogram (ECG) and chest impedance (CI) with three wired adhesive electrodes. CI measures variation in electrical impedance across the chest during respiration caused by changes in lung aeration and chest wall movement. These techniques provide continuous monitoring of heart rate (HR), respiratory rate (RR), and breathing pattern. However, as CI measures respiration indirectly, adequate detection of breathing cycles and apnea may not always be optimal.(3)

With transcutaneous electromyography of the diaphragm (dEMG) breathing effort can be recorded directly by measuring the electrical activity of this main respiratory muscle. To date, this technique also uses three adhesive electrodes and provides information on respiration and HR. Studies have shown its feasibility in the NICU-setting.(4)

The use and especially removal of adhesive electrodes can lead to epidermal stripping in vulnerable preterm infants, resulting in an increased risk of infection and pain.(5, 6) Furthermore, the wires attached to the electrodes restrict movements of the infant and may hinder both nursing and kangaroo care. Restrictions in kangaroo care may impact patient outcome as it has been associated with beneficial effects such as decreased mortality, decreased risk of severe infection/sepsis and hypothermia, and increased likelihood of exclusive breast feeding.(7, 8) Therefore, it is important to find alternatives for using wired adhesive electrodes.

In the past years, several wireless wearable sensors have been developed to measure various parameters in neonates such as ECG, HR, RR, peripheral oxygen saturation and (skin) temperature.(9-15) Recently, a novel wireless and non-adhesive sensor belt (Bambi® belt, Bambi B.V., Eindhoven, the Netherlands) was developed for neonatal use that measures ECG and respiration based on the dEMG technique. A recent pilot study showed that measuring HR and RR with this belt in preterm infants is feasible and that the measured HR and RR trend was similar to ECG and CI.(submitted for publication, NICU AmsterdamUMC, 2021) However, before replacing the current techniques using adhesive wired electrodes with the non-adhesive sensor belt, a larger study is required to demonstrate the non-inferiority of this belt as an alternative cardiorespiratory monitor. In this study, we compare the monitoring performance of the Bambi® belt to ECG and CI and hypothesize that the performance of the belt is non-inferior to the current monitoring techniques.

METHODS

Study design

This multi-center paired non-inferiority study will be performed in the NICU of Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital of the Amsterdam University Medical Centre (AmsterdamUMC), both located in the Netherlands. Each patient will be simultaneously measured with the belt and ECG/CI (paired design). To compare the devices, a non-inferiority/equivalence framework will be used. Here, equivalence is defined as the limit of agreement of the HR/RR between the belt and ECG/CI being within prespecified clinically accepted margins. Non-inferiority is defined as the performance of clinical event detection and quality criteria not being worse than prespecified clinically accepted margins.

Study population

Preterm and term infants being routinely monitored with the standard cardiorespiratory monitor (Intellivue MP90, Philips Healthcare, Eindhoven, The Netherlands) are included in the study. To ensure a representative sample of the target population, infants in different age groups will be included. Infants with chest skin lesions, congenital anomalies, and other scenario's preventing belt placement, such as (effects of) surgery or wrap for therapeutic hypothermia, will be excluded.

Primary outcome

As HR-monitoring is clinically most relied upon and both ECG and dEMG provide the HR by measuring cardiac electrical activity, while CI and dEMG measure respiration with a different technique, the HR is considered the primary outcome.(3, 16) This will be studied with three criteria. 1) Reliable monitoring performance through second-to-second HR measurement agreement between the belt and the ECG-CI monitoring. 2) The detection of a composite cardiac event consisting of bradycardia (HR < 100 beats per minute for at least five seconds)(17) and tachycardia (HR > 180 beats per minute for at least ten seconds)(18) between the belt and the ECG measured with adhesive electrodes. 3) Non-inferior quality (percentage of time with HR recordings without data loss).

Moreover, we will perform subgroup analyses to investigate whether the HR measurement performance is consistent under different clinical activities (e.g. kangaroo care, feeding) and in the different age groups.

Secondary outcomes

The secondary outcome is the measured RR. This will be studied using the following three criteria:

1. Comparing the trend (10-minute moving average) in RR values provided by the belt and CI, as the trend is the primary clinical usage of respiratory monitoring.(3) Since CI is widely used for neonatal respiratory monitoring, it is used as the reference technique.
2. Comparing the ability to detect apnea and tachypnea. Clinically relevant apneas are considered when indicated by a RR < 20 breaths per minute (to capture all periods of low breathing frequency) measured with CI for at least 10 seconds, associated with a desaturation (oxygen saturation (SpO_2) <80% for at least 10 seconds) and/or bradycardia (HR <100 beats per minute for at least five seconds) (objective apnea measurement).(17)

   Tachypnea is defined as a prolonged period of RR >60 breaths per minute and >100 breaths per minute (approximately two times the average normal RR).(19) To cover short and long periods of tachypnea, 3 different durations are studied (30 seconds, 60 seconds, and 10 minutes).
3. Calculating the percentage of time with reliable respiratory monitoring (without data loss and with an acceptable signal-to-noise ratio).

Data collection

The following basic characteristics and demographic information will be collected at the baseline of the study: gestational age, birth weight, gender, age and weight at day of measurement, relevant medical status (respiratory support, medication and underlying illness during measurement), chest circumference, nipple distance, skin type at study start by visual inspection (normal, dry, flaky, oily, moist, other).

Sample size calculation

A power calculation is performed for the primary outcome using data collected in a previous study.(submitted for publication, NICU AmsterdamUMC, 2021) Among the three criteria,

criteria 1 needs the largest sample size and is used for our study. This resulted in 39 required infants to achieve 80% power with an overall 5% type I error with a Bonferroni correction (details in the Statistical Analysis Plan (SAP) in the online supplement).

In addition, an interim analysis will be performed as the power calculation was based on the previous study and recruitment of infants without being able to answer research questions is unethical.(20) This will be performed after including 1/3th of the infants for sample size adaption using the method of Mehta and Pocock.(21) If the conditional power falls within the pre-defined "promising zone", the sample size will be increased to an upper limit of 52 infants. Otherwise, the study will proceed with the original sample size.

Study procedures

The Bambi® Belt System is a non CE-certified medical device, designed for wireless cardiorespiratory monitoring of (pre)term infants in a hospital environment. All included infants will be monitored with the belt in addition to standard ECG/CI for 24 hours to obtain representative clinical scenarios throughout the entire day. The measurement set-up is visualised in Figure 1 and consists of 1) dEMG measurement with the belt and 2) the extraction of patient monitor data.

In the belt, three dry electrodes are incorporated. When placing the belt at the height of the diaphragm, the outer two electrodes are in the nipple line and the middle electrode is in line with the sternum. The three ECG/CI electrodes are attached at the original location without hindering belt placement. The measured electrical signal of the diaphragm with the belt is wirelessly transmitted to the Receiver Module (REM) by the Sensor Module (SEM). The REM processes the dEMG signal to obtain the ECG and respiration signal (averaged diaphragmatic activity). An inbuilt algorithm provides the HR and RR out of the ECG and respiration signal respectively. This data is transported to a bedside computer. The data from the patient monitor

(ECG, HR, RR, and SpO2) is extracted from the bedside monitor using an isolated cable and is also transported to the bedside computer.

The belt data from the REM and patient monitor are recorded and synchronised using a dedicated software package (Polybench, Applied Biosignals, Weener, Germany) on a personal bedside computer. Data is recorded at a sample rate of 1 to 500 Hz for rate and waveform data respectively. The bedside software also provides the possibility to make measurements annotations during data recording, such as re-positioning of the infant, nursing and kangaroo care.

During the study, daily routine care proceeds as usual. The location of the belt is regularly checked and if necessary repositioned (similar to the clinical practice). Notifications are visualised when contact between skin and the belt is lost (Leads off) or when there is no connection between the SEM and REM (Bluetooth Loss Error). In case of the first notification, the belt may be repositioned, while in case of Bluetooth loss the battery level of the SEM or blocking of this sensor (e.g. by an arm) are checked.

Preferably, the belt stays in place during the study. However, the belt can be removed during diagnostic imaging, patient handling, or in case of skin irritation at the belt location. The reason for removal will be annotated. If the belt is removed, the medical staff, parents and one of the dedicated researchers will decide together if the belt can be re-applied.

Recruitment

Parents of all eligible infants are approached for consent to obtain a sample as heterogeneous and representative as possible. Preferably, infants are included as soon as possible after birth.

During the 24 hours, the study can be terminated if requested by parents or the treating physicians. In case of withdrawal of a subject, an extra subject will be included.

Safety

Being a medical device study, this study was classified as a moderate risk.(22) A specified monitor plan for the study is made based on risk-classification.

STATISTICAL ANALYSIS

A detailed SAP can be found in the online supplement. Unless otherwise specified, all hypothesis tests are two-sided with a significance level of 0.05. All statistical analyses will be performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

The non-inferiority/equivalence margins based on expert opinions (survey send to neonatologists of different NICU's in the Netherlands) and literature (4, 23, 24) are described in Table 1. In the different subparagraphs we refer to this table.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***acceptance margins*[#]**
LOA of second-to-second HR differences	± 8 bpm
LOA of RR trend differences	± 15 brpm
Sensitivity of brady-/tachycardia detection	90%
PPV of brady-/tachycardia detection	90%
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data without "Leads off" or "Bluetooth Loss Error" in the belt.

[#] The acceptance margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac interference, all values for PPV for apnea/tachypnea are acceptable. Interpretations will be made based on the results.

Summary and descriptive statistics

Categorical data will be summarized by numbers of counts and percentages. Continuous data will be summarized by mean, standard deviation if data is normal and median, interquartile range (IQR) if data is skewed. Minimum and maximum values will also be presented for continuous data when appropriate.

Statistical analysis of the primary outcome

Criteria 1: agreement in HR

To investigate the equivalence of HR measurement between the belt and ECG, we will fit a linear mixed model to the second-to-second HR difference between both. With this model, the 95% limits of agreement (Bland-Altman analysis) will be derived. The two-one-sided tests (TOST) with a multiplicity corrected alpha of 0.0167 and the prespecified margin ($\pm$ 8bpm) will test equivalence between the two devices. In addition, based on a bivariate heteroscedastic model fitted to HR segments of a prespecified length, additional performance measures will be calculated as sensitivity analyses (details in SAP).

Criteria 2: cardiac event detection

For HR monitoring, we also consider the detection of bradycardia and tachycardia. We will estimate the sensitivity and the positive predictive value (PPV) of the belt using the patient monitor data as the ground truth and perform a non-inferiority test with an alpha of 0.0167. The non-inferiority margin for the sensitivity and PPV are listed in Table 1. In case of missed bradycardias, one independent expert per center will qualify the severity and acceptability of each missing event with the use of the discrepancy in HR and contextual information.

Criteria 3: signal quality

The quality of the investigational device will be quantified based on the percentage of time during the 24-hour period it produces any reading (percentage without data loss due to "Leads off" or "Bluetooth Loss Error") and the percentage in time it produces a good-quality-reading (percentage of robust data) for the HR and RR, respectively. For the HR non-robust data can be caused by bad connection (suboptimal Bluetooth or skin-electrode connection). These criteria are built-in in the belt algorithm and therefore this data is automatically labeled. Hypothesis testing will be used to establish the non-inferiority of this "uptime" percentage (percentage without data loss and percentage of robust data) of the belt.

For the RR, the uptime percentage is also categorized as a) data readings without data loss and b) robust data readings, i.e. good quality readings with acceptable signal-to-noise ratio's undisturbed by among others suboptimal connection and patient handling. Again, the prespecified margins for the HR and RR are described in Table 1.

Statistical analysis of secondary outcomes

Secondary analyses, based on the same statistical methods for the criteria of the primary outcome, include all secondary endpoints (apnea and tachypnea detection, RR trend analysis (see SAP)) and evaluation during different scenarios.

ETHICS AND DISSEMINATION

The Medical Ethical Committee of the MMC (W21.042) and the Central Committee for Human Research in the Netherlands (CCMO, CCMO21/0167/PP) approved the study protocol (Version 2, 19th of May 2021). Local feasibility at the AmsterdamUMC was approved by the Medical Ethical Committee of the AMC (2021_146**).** This study was registered in the Dutch Trial Register (https://www.trialregister.nl, NL9480). Regarding patient safety, no belt related events were observed in the pilot study and are therefore unexpected. Moreover, as every patient is

monitored with ECG/CI and the belt, safety is guaranteed in case of missing belt data. The SAP will be used for the analyses. The results will be published in peer-reviewed journals and presented at future congresses.

The MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The duration of this study will be approximately seven months.

PATIENT AND PUBLIC INVOLVEMENT STATEMENT

Patients were included in this study after obtaining parental informed consent. The patients could not be involved in the design, recruitment, conduction and dissemination of results of this study. Neither could we ask the burden of the study. The outcome measures were developed by combining clinical and statistical knowledge to ensure a SAP that enables confirmation of non-inferiority of the belt compared to ECG/CI.

REFERENCES

1. Joshi R, van Straaten H, van de Mortel H, et al. Does the architectural layout of a NICU affect alarm pressure? A comparative clinical audit of a single-family room and an open bay area NICU using a retrospective study design. *BMJ Open* 2018;8(6):e022813. doi: 10.1136/bmjopen-2018-022813

2. Bouwstra S, Chen W, Feijs LMG, et al. Designing the parents-to-infant bonding experience in the NICU. *IARIA* 2012:7-13.

3. Di Fiore JM. Neonatal cardiorespiratory monitoring techniques. *Semin Neonatol* 2004;9(3):195-203. doi: 10.1016/j.siny.2003.11.009

4. Kraaijenga JV, Hutten GJ, de Jongh FH, et al. Transcutaneous electromyography of the diaphragm: A cardio-respiratory monitor for preterm infants. *Pediatr Pulmonol* 2015;50(9):889-95. doi: 10.1002/ppul.23116

5. Csoma ZR, Meszes A, Abraham R, et al. Iatrogenic Skin Disorders and Related Factors in Newborn Infants. *Pediatr Dermatol* 2016;33(5):543-8. doi: 10.1111/pde.12960

6. Sekar KC. Iatrogenic complications in the neonatal intensive care unit. *J Perinatol* 2010;30 (Suppl):P51-6. doi: 10.1038/jp.2010.102

7. Boundy EO, Dastjerdi R, Spiegelman D, et al. Kangaroo Mother Care and Neonatal Outcomes: A Meta-analysis. *Pediatrics* 2016;137(1). doi: 10.1542/peds.2015-2238

8. Conde-Agudelo A, Diaz-Rossello JL. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *Cochrane Database Syst Rev* 2016(8). doi: https://doi.org/10.1002/14651858.CD002771.pub4

9. Chung HU, Rwei AY, Hourlier-Fargette A, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. *Nat Med* 2020;26(3):418-29. doi: https://doi.org/10.1038/s41591-020-0792-9

10. Henry C, Shipley L, Ward C, et al. Accurate neonatal heart rate monitoring using a new wireless, cap mounted device. *Acta Paediatr* 2021;110(1):72-8. doi: https://doi.org/10.1111/apa.15303

11. Rimet Y, Brusquet Y, Ronayette D, et al. Surveillance of infants at risk of apparent life threatening events (ALTE) with the BBA bootee: a wearable multiparameter monitor. *Annu Int Conf IEEE Eng Med Biol Soc* 2007;2007:4997-5000. doi: 10.1109/IEMBS.2007.4353462

12. Chen H, Bao S, Lu C, et al. Design of an Integrated Wearable Multi-Sensor Platform Based on Flexible Materials for Neonatal Monitoring. *IEEE* 2020:23732-47. doi: 10.1109/ACCESS.2020.2970469

13. Chen W, Dols S, Bambang Oetomo S, et al. Monitoring body temperature of newborn infants at neonatal intensive care units using wearable sensors. *BodyNets '10* 2010. doi: https://doi.org/10.1145/2221924.2221960

14. Linti C, Horter H, Osterreicher P, et al. Sensory baby vest for the monitoring of infants. *IEEE* 2006. doi: 10.1109/BSN.2006.49

15. Pike H, Eilevstjonn J, Bjorland P, et al. Heart rate detection properties of dry-electrode ECG compared to conventional 3-lead gel-electrode ECG in newborns. *BMC Res Notes* 2021;14(1):166. doi: 10.1186/s13104-021-05576-x

16. van Leuteren RW, Hutten GJ, de Waal CG, et al. Processing transcutaneous electromyography measurements of respiratory muscles, a review of analysis techniques. *Journal of Electromyography and Kinesiology* 2019:176-86. doi: 10.1016/j.jelekin.2019.07.014

17. Lee H, Rusin CG, Lake DE, et al. A new algorithm for detecting central apnea in neonates. *Physiol Meas* 2012;33(1):1-17. doi: 10.1088/0967-3334/33/1/1

18. Li T, Matsushima M, Timpson W, et al. Epidemiology of patient monitoring alarms in the neonatal intensive care unit. *J Perinatol* 2018;38(8):1030-8. doi: 10.1038/s41372-018-0095-x

19. Reuter S, Moser C, Baack M. Respiratory distress in the newborn. *Pediatr Rev* 2014;35(10):417-28. doi: 10.1542/pir.35-10-417

20. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

21. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

22. Central Committee on Research Involving Human Subjects.

23. Anton O, Fernandez R, Rendon-Morales E, et al. Heart Rate Monitoring in Newborn Babies: A Systematic Review. *Neonatology* 2019;116(3):199-210. doi: 10.1159/000499675

24. Breteler MJM, KleinJan EJ, Dohmen DAJ, et al. Vital Signs Monitoring with Wearable Sensors in High-risk Surgical Patients: A Clinical Validation Study. *Anesthesiology* 2020;132(3):424-39. doi: 10.1097/ALN.0000000000003029

AUTHOR CONTRIBUTIONS

AS, HN, MV, RL, FJ, AK, JH conceptualized the study. ZZ and EH made the statistical analysis plan, which was reviewed by all authors. AS wrote the first version of this manuscript. All authors contributed to the final draft of the manuscript.

FUNDING STATEMENT

This work was supported by the Louise Vehmeijer foundation.

COMPETING INTERESTS STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

WHAT IS KNOWN ABOUT THE SUBJECT:

- Disadvantages of the cardiorespiratory monitoring technique in neonates are indirect measurements of respiration, usage of adhesive electrodes and hindering wires.
- With transcutaneous electromyography of the diaphragm, respiratory activity is measured directly by recording the activity of the main respiratory muscle.
- The Bambi® belt is a novel wireless and non-adhesive belt that enables cardiorespiratory monitoring by measuring diaphragm activity with dry electrodes.

WHAT THIS STUDY HOPES TO ADD:

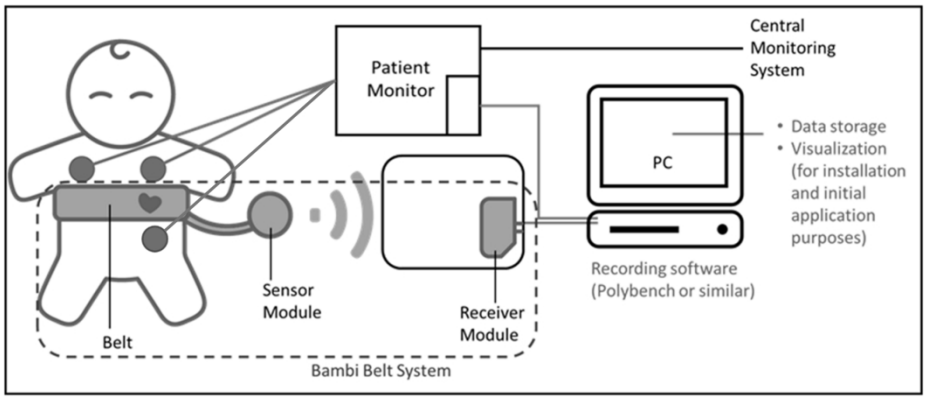
- Demonstration of the non-inferiority of the Bambi® belt compared to the electrocardiogram and chest impedance for cardiorespiratory monitoring in preterm and term infants.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- When non-inferiority of the Bambi® belt compared to the current cardiorespiratory monitor is confirmed, the belt could be used as a wireless and skin-friendly alternative.

FIGURES

FIGURE 1 - The measurement set-up. The adhesive electrodes used for standard cardiorespiratory monitoring are attached at the original location, visualised by the three blue dots. The diaphragm activity measured with the Bambi® belt is wirelessly transmitted with the Sensor Module to the Receiver Module where the data is processed to obtain an electrocardiogram and respiration waveform (and heart rate and respiratory rate). This data and the data measured with the patient monitor are transported to a personal bedside computer with Polybench software to synchronise and record these signals.



170x80mm (300 x 300 DPI)

ONLINE SUPPLEMENT – STATISTICAL ANALYSIS PLAN

PURPOSE OF THE STATISTICAL ANALYSIS PLAN

The purpose of this statistical analysis plan (SAP) is to document technical and detailed specification for the analysis of data collected in the Bambi Belt monitoring performance study. The SAP has been written based on information contained in study protocol, dated 12th April 2021 before any data collection had taken place. It is prepared in compliance with the International Council on Harmonization (ICH) E9.

This SAP will be the guiding document for the analyses that will be conducted. Results of the analyses described in this SAP will be included in the Clinical Study Report (CSR). Any post hoc or unplanned analyses performed to provide results for inclusion in the CSR, but not identified in the prospective SAP will be identified in the given report. Additionally, the planned analyses of the primary aims will be included in future manuscripts. All the aims and research questions will be presented as an addendum as well.

OVERVIEW AND DESCRIPTION OF THE STUDY

Study design

The study is a multicenter, paired design, clinical monitoring device measurement comparison study. The *investigational device* under consideration is the Bambi® Belt monitoring system (using dry electrodes). The current standard device of cardiorespiratory monitoring through adhesive electrodes is considered as the clinical reference standard and thereafter referred to as the *reference device/method*. The Bambi Belt® monitoring system will be used on infants by trained nurses in the neonatal intensive care units (NICU's) for continuous 24 hours monitoring in addition to the routine monitoring with the reference device on the same patients. Infants admitted to NICU's of the the Emma Children's Hospital of the Amsterdam University Medical

Centre (Amsterdam UMC) or Maxima Medical Center (MMC) will be measured at the earliest suitable moment for clinical practice without interfering with infants' routine cycles.

Randomization and blinding

No randomization is required for the paired design since both monitoring devices will be used on the same patient at the same time. Blinding is also not possible since both the measurement protocol and algorithmic characteristic differ substantially.

Framework

The goal of this study is to establish the agreement between the investigational device and the reference device. Unlike the traditional difference-based tests, non-inferiority and equivalence techniques provide a better alternative for demonstrating the similarity between the two measurement methods. Thus, we have adopted the non-inferiority/equivalence trial framework for this primary objective of this study. This study considers three hypotheses (H_0 denotes the null hypothesis and H_A denotes the alternative hypothesis) for the first two primary outcomes:

1. Primary outcome, criterion 1: Heart rate measurement (second-by-second measurement)

H_0: The absolute difference between the investigational device and the reference device is larger than the prespecified equivalence margin.

H_A: The absolute difference between the investigational device and the reference device is within the prespecified equivalence margin.

2. Primary outcome, criterion 2: Brady-/tachy-cardia event detection

H_0: The composite cardiac event detection performances in terms of sensitivity and positive predictive value (PPV) based on the investigational device with respect to the reference device is less than the prespecified non-inferiority margin.

H_A: The composite cardiac event detection performances in terms of sensitivity and PPV based on the investigational device with respect to the reference device is greater or equal to the prespecified non-inferiority margin.

3. Primary outcome, criterion 3: Reliable reading (percentage of the time)

H_0: The percentage of the time the investigational device produces reliable readings is less than the prespecified non-inferiority margin.

H_A: The percentage of the time the investigational device produces reliable readings is greater or equal to the prespecified non-inferiority margin.

Statistical interim analysis and stopping guidance

One interim analysis for sample size adaptation will be performed. That is, we will start with a certain sample size commitment which will be increased at the interim analysis in case the results obtained are reasonably promising. The interim analysis will be conducted after the prospectively recruited participant's number reaches one-third of the planned sample size. Conditional power will be calculated for the analyses of the primary endpoints and compared to the boundary values of the conditional power for the promising zones (1, 2). In case the conditional power calculated at the interim analysis does fall inside the promising zone, the sample size will be increased to a pre-determined limit. On the other hand, if the calculated conditional power is outside the promising zone, the study will proceed with the original sample size. Therefore, no early stopping rule is entailed in this study. Furthermore, a conventional

final analysis will be used without altering the level of type I error, since the promising zone is defined as a set that ensures the type I error to be preserved conservatively for the final analysis.

Study data

The following patient characteristics will be collected at baseline:

- Gestational age
- Postmenstrual age
- Gender
- Weight at enrollment
- Ethnicity
- Chest circumference
- Nipple distance
- Skin condition and abnormality

During the monitor study period, the following information will be measured:

- Clinical event
- SpO2: Blood oxygen saturation level
- Medical status:
    - Ventilation support
    - Reports of medication and illness during the measurement
- Lead status: Indicates whether at least one lead was off
- Bluetooth link quality
- Activities
    - Kangaroo care
    - Nurse care
    - Feeding

- Medical Procedure

- Belt status
  - Moved: the belt is being moved
  - Open: the belt is removed from the patient
- Patient position
  - Unknown
  - Lying prone
  - Lying supine
  - Lying on the left side
  - Lying on the right side

STATISTICAL ANALYSIS

Based on the collected information described above, the following total of variables will be derived:

- 10, 30, and 60 minutes moving average of the heart rate, and respiratory rate measured by both the investigational device and the reference device.
- Premature birth:
  - Premature (gestational age < 37 weeks)
  - Normal (gestational age >= 37 weeks)
- Desaturation: SpO2 < 80% for at least 10 consecutive seconds
- Heart rate status (investigational and reference device):
  - Normal
  - Tachycardia (heart rate > 180 for at least 10 consecutive seconds)
  - Bradycardia (heart rate < 100 for at least 5 consecutive seconds)
- Respiration status (investigational and reference device):

- Apnea (according to standard clinical definitions)
- Tachypnea (respiratory rate >60 and >100 for 30 seconds, 1 minute, and 10 consecutive minutes in stationary signal)

- Measurement quality:
  - No anomalies
  - Poor data link: Bluetooth link is poor but data is still received
  - Unreliable data: One or more lead off, or no Bluetooth connection (Bluetooth Loss Error, BLE)

Summary and descriptive statistics

Categorical data will be summarized by numbers and percentages. Continuous data will be summarized by mean, standard deviation if data are normal and median, interquartile range (IQR) if data are skewed. Minimum and maximum values will also be presented for continuous data. Tests of statistical significance will not be undertaken for baseline characteristics; rather the clinical importance of any imbalance will be noted.

A CONSORT flow diagram (example in Figure 1) will be used to summarize the number of infants who were:

- Assessed for eligibility at the screening
  - Eligible at screening
  - Ineligible at screening (with reasons)
- Eligible and enrolled
- Eligible but not enrolled
- Enrolled but did not receive any / sufficient measurements
  - Discontinued

- Included in the analysis
- Excluded in the analysis

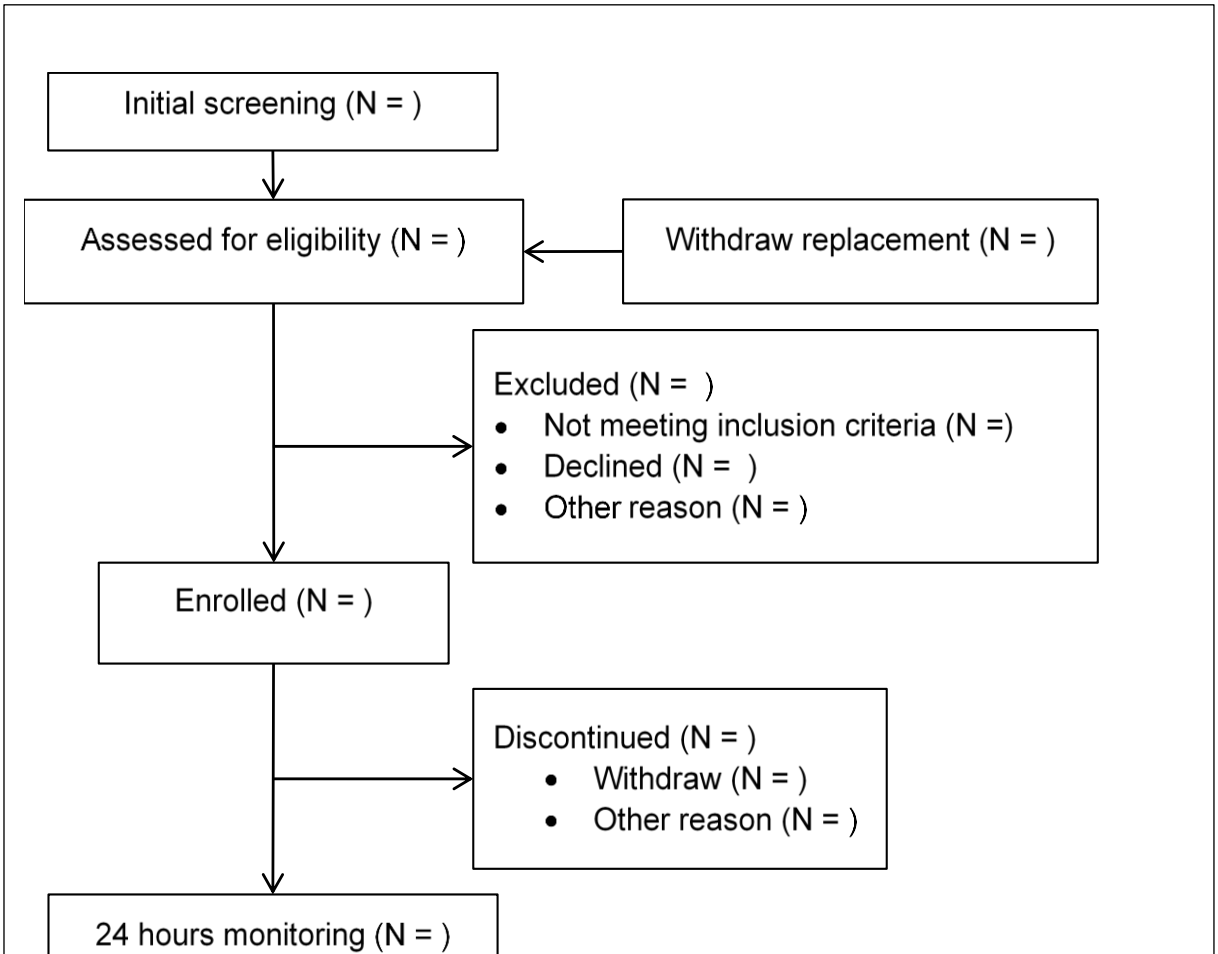


FIGURE 1 – Example of a CONSORT flow diagram

Analysis methods

Primary outcome, criterion 1: Heart rate measurement

To investigate and verify the equivalence of heart rate measurement between the investigational device and the reference device, we will fit a linear mixed model to the second-to-second heart rate difference between the two devices. Based on the estimates of the model, we will derive the 95% limits of agreement (3) as our main performance measure, known as the Bland-Alman analysis. The endpoints of the Bland-Altman 95% limits of agreement are the 2.5th percentile and 97.5th percentile for the distribution of the difference between paired measurements. We will calculate the $(1-\alpha/2)100\%$ confidence intervals of the percentiles according to Shieh (4), and conduct the two-one-sided t-tests (TOST) procedure with the prespecified equivalence margins (Table 1).

In addition, we will calculate the following performance measures to supplement the main analysis as sensitivity analyses to assess the agreement between the two devices from different aspects:

- The concordance correlation coefficient (5) and its variants
- Probability of Agreement (6) and Total Deviation Index (7)
- Coefficient of individual agreement (8)

These performance measures will be based on a bivariate heteroscedastic linear mixed-effects model fitted to each segment of the readings of a prespecified length from both devices. We will assume that measurements made between the two devices at the same time are correlated. Therefore, investigating the correlation between the two devices leads to the quantification of the degrees of agreement between them. Furthermore, we will consider the temporal correlations between measurements obtained with the same devices and the variabilities between different infants. Besides, we will start with a heteroscedastic model which does not assume equal variances for the two devices (namely, the measurement errors are not assumed

to be equal) and investigate the homogeneity of the measurement variabilities between the two devices. Baseline characteristics of the infants and records of activities (listed in the study data section) will be used as covariates in the model to partly explain the variabilities between the infants. We will use the stepwise model selection procedure based on the Bayesian Information Criteria (BIC) goodness-of-fit criteria.

Primary outcome, criterion 2: Brady-/tachycardia event detection

For brady-/tachycardia, the clinical event periods will be identified based on pre-specified clinical definitions. We will investigate the non-inferiority of sensitivity and positive predictive values (PPV) of the event detected by the investigation device assuming that the reference device is the predicate device and compare both values to the prespecified non-inferiority margins (Table 1). For the calculation of the sensitivity, when the event period identified based on the investigational device overlaps with the event period identified by the reference device, it will be counted as a true positive case. This is to prevent the repeated signaling of events from the investigational device during a positive period identified by the reference device to inflate the number of true positives. The same applies to the reference device when it comes to the calculation of the PPV. That is, during an event period identified by the investigational device, multiple event periods identified by the reference device will only be counted as one true-positive case. Note that the true negative is ill-defined and will not be reported. Since true negatives are used in the calculation of specificity, specificity will not be reported either.

Primary outcome, criterion 3: Safety and Quality

Safety: The investigation of safety and tolerability is a multidimensional problem. Although we don't anticipate any specific adverse effects for the investigational device, new and unforeseeable effects are always possible. This background underlies the statistical difficulties

associated with the analytical evaluation of the safety and tolerability of the device. We will address the safety and tolerability implications by applying descriptive statistical methods to the data, supplemented by calculation of confidence intervals whenever this aids interpretation and make use of graphical presentations in which patterns of adverse events are displayed.

Quality: The quality of the investigational device will be quantified in terms of the point estimate and 95% confidence intervals based on the estimated percentages in time during the 24-hour period it produces reliable readings for heart rate and respiratory rate, respectively. Reliable readings are defined in the study protocol. The uptime percentages are the percentage of data loss and the percentage of robust data readings. For each outcome, hypothesis testing will be used to establish the non-inferiority of the uptime percentages of the investigational device considering a non-inferiority margin specified in Table 1. The uptime percentages will be estimated based on a Generalized Estimation Equations (GEE) model.

Missing data

To get an idea about the complexity of the missing data problem in the data and information about the location of the missing values, the missing data pattern will be evaluated and reported. We expect missing data in the primary outcomes measured by the investigational device to be the results of external causes such as the movement of the belt, signal losses, poor Bluetooth link qualities and so on. Therefore, it will be reasonable to assume that data are missing completely at random (MCAR). Formally, we will investigate the validity of such an assumption using Little's MCAR test. Furthermore, the availability of the data from the reference device (since it depends on a separate measurement system) provides us the opportunity to investigate whether the missingness is related to the underlying measurand. That is, whether the missing data mechanism is missing not at random (MNAR). This is rarely

possible in other types of studies. Nevertheless, considering the pair of bivariate measurements from the investigational and the reference device, we will investigate the assumption using the covariate-dependent missing (CDM) test proposed in Li (9). Note that CDM is usually considered as missing at random (MAR), we here simply exploit the advantage of the data from the reference device to test the dependencies between the missingness and the underlying measurand. Furthermore, we will use the CDM test on other covariates (excluding the reference device data) as well to test if the missingness is MAR.

In the case of MAR (i.e., CDM without measurements from reference device), list wise deletion can still be unbiased and will be used if the percentage of missingness is less than 5%. Otherwise, multiple imputations (MI) will be considered. We will not use the measurements from the reference device for the MI to avoid biasing the results towards the equivalence of the two devices. On the other hand, if the missingness is related to the measurand after taking into account all covariates, this indicates a potential problem of the measurement device, and a separate analysis will be carried out to investigate the associations between the missingness and the measurand.

For multiple imputations, we will use the fully conditional specification method. Unrealistic values (e.g., negative values for strictly positive variable) will be checked and corrected (e.g., using truncations). The imputation will be repeated at least 5 times and Rubin's rule will be used to combine estimates and standard errors from the imputed data.

Secondary analyses

If the sample size permits, we will perform subset analyses to explore the performances of the investigational device under different scenarios.

Subset analyses: primary endpoints

For each of the primary endpoints, we will consider additional exploratory analyses on the following subsets:

- During periods of a clinical event (e.g., apnea, bradycardia)
- During activities (e.g., Kangaroo care, feeding)
- During periods where the reference device's readings are stable
- Gestational age (e.g., preterm birth)
- Respiratory support (e.g., mechanical ventilation)

For these subsets, we will use the same model as the primary outcome to investigate the performances of the investigational device under various scenarios/activities of the infants. In case the subset does not contain enough data to fit the same model as the primary one, we will resort to a simpler model for case-by-case analyses.

Respiratory rate analysis

It is known the reference device does not provide point-by-point accurate measurement resulting in large variabilities (measurement errors) in the measured respiratory rates. The intended clinical use of the readings in the NICU thus consists of two different aspects:

1. The trend of the respiratory rates over time;
2. Signaling of potentially respiratory related clinical events (i.e. apnea related desaturation and/or bradycardia, and potentially disease related tachypnea);

For the first usage, we will apply the same analysis method as the one used for heart rate on the moving average of the respiratory rate. We will primarily focus on the 10 minutes moving average for the respiratory rate. Analysis of the 1 minute and 5 minutes moving averages will

be used as a sensitivity analysis to establish the robustness of the conclusions made for the 10 minutes moving average.

For apnea and tachypnea, respectively, the clinical event periods will be identified based on clinical definitions and the same methods as the brady-/tachycardia event detection will be used to compare the sensitivity and PPV to the prespecified non-inferiority limits (Table 1). However, it should be noted that since the reference device is known to have an unsatisfactory performance of apnea/tachypnea detection, cautions are needed to interpret the sensitivity and PPV as if the reference device is the truth.

Statistical software

All statistical analyses will be performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

Non-inferiority/equivalence criteria

In Table 1 the non-inferiority/equivalence criteria for the primary and secondary outcomes are visualized.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***acceptance margins#***
LOA of second-to-second HR differences	±8 bpm
LOA of RR trend differences	±15 brpm
Sensitivity of brady-/tachycardia detection	90%
PPV of brady-/tachycardia detection	90%
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data without "Leads off" or "Bluetooth Loss Error" in the belt.

The acceptance margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac interference, all values for PPV for apnea/tachypnea are acceptable. Interpretations will be made

Sample size

Based on preliminary analysis of data collected in a feasibility study on a total of 13 infants with measurements from both the investigational device and the reference device, we were able to obtain preliminary information with regards to the characteristics of the primary endpoints

upon which we have formulated our sample size calculation.(submitted for publication, NICU AmsterdamUMC, 2021)

A detailed specification of the sample size calculation can be found in the sections below. In summary, for the monitor performance, 39 infants are needed to achieve 80% power with a 5% overall type I error with a Bonferroni correction for multiplicity. It is worth noting that no dropout was assumed during the sample size calculation. This is because we plan to include an extra infant in case of withdrawal of an infant to fulfil the required sample size. Infants who withdraw from the study will be followed up by one of the investigators and responsible medical staff to obtain detailed reasons behind the withdraw. Dropout rate for the monitor performance study is expected to be low, between 0-5%.

While the pre-planned sample size is 39 infants, we will include an adaptive sample size re-estimation procedure as per the "promising zone" methodology of Mehta and Pocock (2) using the data from the first 1/3 infants. This procedure involves the evaluation of conditional power in the interim analysis, and if it were to fall in the pre-specified "promising zone", the sample size will be increased, subject to a pre-determined upper limit (52 infants) to increase the conditional power to 80%. The boundary of the conditional power for the "promising zone" is 0.36 and 0.8.

Monitoring study: Primary endpoints: Heart rate

For the sample size calculation, we will assume the measured heart rate difference D_{ij} between the investigational device and the reference device at time point j $(j = 1, \dots, m)$ on infant i $(i = 1, \dots, n)$ can be modelled as:

$$D_{ij} = d + a_i + e_{ij}$$

where d is the overall difference, a_i is a random effect with $a_i \sim N(0, \sigma_a^2)$, and $e_{ij} \sim N(0, \sigma_e^2)$ is the random error independent of a_i. Though, we considered a bivariate mixed-effects model for our analysis, the variance component model for the difference can be derived from the bivariate mixed-effects model, therefore we will use this variance component model for the sample size calculation. The variance of the difference will be estimated from the aforementioned model via $\hat{\sigma}_d^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2$. Here $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ is the estimator of the between-subject variability σ_a^2 and residual variability σ_e^2, respectively. The 95% limit of agreement (LOA) can be estimated as LOA = $\hat{d} \pm 1.96\,\hat{\sigma}_d$ with $\hat{d}$ and $\hat{\sigma}_d$ denotes the estimator of d and σ_d, respectively. The variance of the LOA estimator is $\text{var}\left(\hat{d} \pm 1.96\,\hat{\sigma}_d\right) = \text{var}\left(\hat{d}\right) + 1.96^2\text{var}(\hat{\sigma}_d)$ ($\hat{d}$ and $\hat{\sigma}_d$ is asymptotically independent). Since for $\hat{\sigma}_d^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2$, we have $\text{var}(\hat{\sigma}_d^2) = \text{var}(\hat{\sigma}_a^2) + \text{var}(\hat{\sigma}_e^2) + \text{cov}(\hat{\sigma}_a^2, \hat{\sigma}_e^2)$. Furthermore, each term on the right-hand side (assuming m is large) is given by:

$$var\,(\hat{\sigma}_a^2) = \frac{2}{m^2}\left[\frac{(m\sigma_a^2 + \sigma_e^2)^2}{n-1} + \frac{\sigma_e^4}{n(m-1)}\right] \approx \frac{2\sigma_a^4}{n-1}, var\,(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{n(m-1)+2} \approx 0,$$

$$cov(\hat{\sigma}_a^2, \hat{\sigma}_e^2) = -\frac{2\sigma_e^4}{nm(m-1)} \approx 0;$$

This leads to $\text{var}\left(\hat{\sigma}_d{}^2\right) \approx 2\sigma_a{}^2/(n-1)$. Therefore, by the delta method, we have $\text{var}(\hat{\sigma}_d) = \frac{1}{4\sigma_d{}^2} var(\hat{\sigma}_d{}^2) = \frac{\sigma_a{}^4}{2(n-1)\sigma_d{}^2}$. According to Lu et al. (10), the power for the TOST is given by:

$$1 - \beta = 1 - T_{n-1}\left(t_{1-\frac{\alpha}{2}}, \frac{\delta - d - 1.96\sigma_d}{se_{LOA}}\right) - T_{n-1}\left(t_{1-\frac{\alpha}{2}}, \frac{\delta + d - 1.96\sigma_d}{se_{LOA}}\right)$$

where α, β denotes type I and type II error respectively, δ is the pre-defined acceptance limit that is clinically acceptable, $se_{LOA} \approx \sqrt{\frac{{\sigma_d}^2}{n} + \frac{1.96^2 {\sigma_a}^4}{2(n-1){\sigma_d}^2}}$ is the standard error of the LOA estimate calculated according to the variance component model, and $T_{n-1}(\cdot, \tau)$ denotes the cumulative distribution function of a non-central Student's t-distribution with n-1 degrees of freedom, and non-centrality parameter τ.

For a 5% overall type I error rate, with a multiplicity correction factor of 3, and 80% power, the minimum sample size required is calculated at $n = 39$, for $d = -0.5$, $\sigma_a = 0.3$, and $\sigma_d = 3$.

Primary endpoints: Brady-/tachycardia event detection

Suppose the total number of true events is M and are 100% detected by the reference device. Assuming the true sensitivity is 95% for the investigational device, then a non-inferiority test using Z-test with normal approximation to the binomial distribution leads to a required M of 271 for a power of 80% and $\alpha = 0.05/3 \approx 0.01667$ assuming the detection between each event (conditioning on the event itself) is independent. Considering the incidence of bradycardia to be 1 event per hour per infant according to the preliminary analysis of data from the feasibility study, at least 12 infants are needed to satisfy the required M (assuming each infant is measured for 24 hours long). The calculation is the same for PPV if we assume the investigational device is the truth. Assuming an incidence rate of 1.5 per hour per infant according to the preliminary analysis of data from the feasibility study, the required sample size is 8. Note that in the aforementioned calculation, we assume that the event-detection

performance of the investigational device is homogeneous (or independent) among infants. A sensitivity/robustness investigation regarding the sample size for infant-specific heterogeneous performances was performed, with results from which we can see that with $n = 39$, we have more than 90% power to detect a heterogeneous performance scenario where 15% of the population would have sensitivities between 80% - 90% and less than 5% of the population have sensitivities less than 80%.

Primary endpoints: Safety and quality

Based on preliminary analysis of data collected in the feasibility study, we will assume that the overall probability of producing an erroneous reading at any time p_e is 2% and is constant across all participants. We will consider a non-inferiority test using normal approximation and a Z-test with the null hypothesis of $H_0: p_e > 0.05$ and the alternative hypothesis of $H_A: p_e \leq 0.05$. The required number of observations for a given type I error of 1.667% ($\approx 5\%/3$) to achieve 80% power is 376. Here the sample size 376 refers to 376 independent observations. Considering the large numbers of repeated measurements (more than 376) within each participant, we will have sufficient power for this non-inferiority test even with 1 participant. However, the assumption of independence can be too strong in the setting of our study. Therefore, if we would assume an AR(1)-type dependency with correlation parameter $\rho = 0.8$ between two measurements within a participant, the variance inflation factor (VIF) for the asymptotic variance of the GEE estimator $\hat{p}_e$ according to Pan is approximately (with the number of repeats $m = 376$):

$$1 + \frac{2\rho}{1-\rho} = 9$$

in the case of identity working correlation matrix when the true correlation has an AR(1) structure. To achieve the same power as the independent case calculated before, we need

$$var(\hat{p}_e) \coloneqq VIF\,\frac{p_e(1-p_e)}{nm} = \frac{p_e(1-p_e)}{m}$$

Thus, we can conclude that at least $n = \text{VIF} = 9$ participants will be needed to provide enough power for the non-inferiority test based on the GEE estimator using the identity working correlation matrix using the inverse proportionality between the required sample size and the variance of the estimator used in the Z-test. The same calculation can be carried for the robust data percentages. It can be seen that only the number of repeats m will differ when the probabilities and the non-inferiority margins change while the VIF remains the same for the same value of the correlation parameter ρ. Among all settings, the largest m needed will be 718 when we assume the probability of producing robust data for respiratory rate is 75% with the corresponding non-inferiority margin equals to 70%. This number of repeats is still fully covered by the high-frequency measurements found in the study.

Protocol deviations and analysis sets

Definition of protocol deviations

Protocol deviations (PD) occurring during the study will be determined for all enrolled infants, mainly from the clinical database by either clinical and/or medical review processes.

The mapping of the protocol deviations from the clinical database to analysis will be performed as per Table 2:

Table 2 – The influence of protocol deviations on the statistical analysis plan (SAP)

Database label	*SAP*
Minor	Not required
Major	Important
Critical	Important
Clinical (a subset of Critical)	Important

Important protocol deviations are protocol deviations that might significantly affect the completeness, accuracy, and/or reliability of the study data or that might significantly affect a subject's rights, safety, or well-being.

Important protocol deviations may also be recorded as "Major" protocol deviations in the database, but will be presented only as important in the analysis output.

Important protocol deviations include:

- Infants that are included in the study despite not satisfying the eligibility criteria;
- Infants that develop exclusion criteria while on the study but not withdrawn;
- Infants being measured with operational human errors;
- Deviation from Good Clinical Practice (ICE E6)

Clinically Important protocol deviations are a subset of important protocol deviations which lead to the exclusion of a subject from the analysis set.

The following deviations will be identified and confirmed before the partial database lock for the final analysis.

- Important protocol deviations including
    - Deviations from the inclusion and exclusion criteria
    - Deviations post inclusion

Protocol deviations may be identified by the data managers, clinical and medical staff either by programmed validation checks or data listings/reports or manual verification of data sources. Some important/major protocol deviation criteria may be identified in the clinical database via biostatistical programs. Every important protocol deviation will be documented in the database whether identified through sites monitoring, medical review or programming.

REFERENCES

1. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

2. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

3. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8(2):135-60. doi: 10.1177/096228029900800204

4. Shieh G. The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. *BMC Med Res Methodol* 2018;18(1):45. doi: https://doi.org/10.1186/s12874-018-0505-y

5. Lin L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989:255-68.

6. Stevens NT, Steiner SH, MacKay RJ. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Stat Methods Med Res* 2017;26(6):2487-504. doi: 10.1177/0962280215601133

7. Lin L. I.-K. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 2000:255-70.

8. Haber M, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Stat Methods Med Res* 2008;17(2):151-69.

9. Li C. Little's Test of Missing Completely at Random. *The Stata Journal* 2013. p. 795-809. doi: 10.1177/1536867X1301300407

10. Lu MJ, Zhong WH, Liu YX, et al. Sample Size for Assessing Agreement between Two Methods of Measurement by Bland-Altman Method. *Int J Biostat* 2016;12(2). doi: 10.1515/ijb-2015-0039

BMJ Paediatrics Open

Multi-center paired non-inferiority study of the cardiorespiratory monitoring performance of the wireless and non-adhesive Bambi® belt measuring diaphragm activity in neonates: Study Protocol

Journal:	*BMJ Paediatrics Open*
Manuscript ID	bmjpo-2022-001430.R1
Article Type:	Protocol
Date Submitted by the Author:	21-Mar-2022
Complete List of Authors:	Scholten, Anouk; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute Zhuozhao, Zhan; Eindhoven University of Technology, Department of Mathematics and Computer Science Niemarkt, Hendrik; Máxima Medisch Centrum, Neonatology Vervoorn, Marieke; Máxima Medisch Centrum, Neonatology van Leuteren, Ruud; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute de Jongh, Frans; Amsterdam UMC Locatie AMC, Neonatology; University of Twente, Faculty of Science and Technology van Kaam, Anton; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie VUmc, Neonatology van den Heuvel, Edwin; Eindhoven University of Technology, Department of Mathematics and Computer Science Hutten, G. Jeroen; Amsterdam UMC Locatie AMC, Neonatology; Amsterdam UMC Locatie VUmc, Neonatology
Keywords:	Neonatology, Technology

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

MULTI-CENTER PAIRED NON-INFERIORITY STUDY OF THE CARDIORESPIRATORY MONITORING PERFORMANCE OF THE WIRELESS AND NON-ADHESIVE BAMBI® BELT MEASURING DIAPHRAGM ACTIVITY IN NEONATES: STUDY PROTOCOL

Anouk W.J. Scholten[1,2*†], Zhan Zhuozhao[3†], Hendrik J. Niemarkt[4], Marieke Vervoorn[4], Ruud W. van Leuteren[1,2], Frans H. de Jongh[1,5], Anton H. van Kaam[1,6], Edwin R. van den Heuvel[3], G. Jeroen Hutten[1,6]

[1]*Department of neonatology, Emma Children's Hospital, AmsterdamUMC, University of Amsterdam, Amsterdam, The Netherlands*

[2]*Amsterdam Reproduction & Development research institute, Amsterdam, The Netherlands*

3Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[4]*Department of neonatology, Màxima Medical Center, Veldhoven, The Netherlands*

[5]*Faculty of Science and Technology, University of Twente, Enschede, The Netherlands*

[6]*Department of neonatology, Emma Children's Hospital, AmsterdamUMC, Vrije Universiteit, Amsterdam, the Netherlands*

† These authors have contributed equally to this work and share first authorship

CORRESPONDING AUTHOR:

Anouk W.J. Scholten

Department of Neonatology

Emma Children's Hospital, AmsterdamUMC, location AMC

Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

Phone: +31 20 5663675

Email: a.w.j.scholten@amsterdamumc.nl

ABSTRACT

Introduction: Cardiorespiratory monitoring is used in the Neonatal Intensive Care Unit (NICU) to assess the clinical status of newborn infants and detect critical deteriorations in cardiorespiratory function. Currently, heart rate is monitored by electrocardiography (ECG) and respiration by chest impedance (CI). Disadvantages of current monitoring techniques are usage of wired adhesive electrodes which may damage the skin and hinder care. The Bambi® belt is a wireless and non-adhesive alternative that enables cardiorespiratory monitoring by measuring electrical activity of the diaphragm via transcutaneous electromyography (dEMG). A previous study showed feasibility of the Bambi belt and this study compares the belt performance to ECG and CI.

Methods and analysis: This multi-center non-inferiority paired study will be performed in the NICU of the Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital AmsterdamUMC in Amsterdam, the Netherlands. 39 infants in different postmenstrual age groups (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks) will be recruited. These infants will be monitored with the Bambi® belt in addition to standard ECG and CI for 24 h. The primary outcome is the heart rate (HR), studied with three criteria: 1) the ~~agreement in~~limits of agreement of the HR measurements in terms of the second-to-second difference in the HR between the belt and standard ECG, 2) the detection of cardiac events consisting of bradycardia and tachycardia and 3) the quality of HR-monitoring. The secondary outcome is the respiratory rate (RR), studied with the criteria 1) agreement in ~~respiratory rate (RR)~~ RR trend monitoring, 2) apnea and tachypnea detection, and 3) reliable registrations.

Ethics and dissemination: This protocol was approved by the Medical Ethical Committee of the Màxima Medical Center and the Central Committee for Human Research (CCMO). The

MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The results will be presented at conferences and published in peer-reviewed journals.

Trial registration number: NL9480 (www.trialregister.nl)

INTRODUCTION

In the Neonatal Intensive Care Unit (NICU), cardiorespiratory monitoring is crucial to assess clinical condition and to timely detect and treat frequently occurring cardio-respiratory events to prevent morbidity and mortality.(1, 2) To date, this is performed by measuring the electrocardiogram (ECG) and chest impedance (CI) with three wired adhesive electrodes. CI measures variation in electrical impedance across the chest during respiration caused by changes in lung aeration and chest wall movement. These techniques provide continuous monitoring of heart rate (HR), respiratory rate (RR), and breathing pattern. However, as CI measures respiration indirectly, adequate detection of breathing cycles and apnea may not always be optimal.(3)

With transcutaneous electromyography of the diaphragm (dEMG) breathing effort can be recorded directly by measuring the electrical activity of this main respiratory muscle. To date, this technique also uses three adhesive electrodes and provides information on respiration and HR. Studies have shown its feasibility in the NICU-setting.(4)

The use ~~and especially removal~~ of adhesive electrodes is restricted in infants with a postmenstrual age <26 weeks in fear of skin damage. ~~electrodes can lead to epidermal stripping in vulnerable preterm infants, resulting in an increased risk of infection and pain~~.(5) Moreover, electrode removal may cause discomfort. Furthermore, the wires attached to the electrodes restrict movements of the infant and may hinder parent-infant interaction, ~~both~~ nursing and

kangaroo care. Restrictions in kangaroo care may impact patient outcome as it has been associated with beneficial effects such as decreased mortality, decreased risk of severe infection/sepsis and hypothermia, and increased likelihood of exclusive breast feeding.(6, 7) All things considered, ~~Therefore,~~ it is important to find alternatives for using wired adhesive electrodes.

In the past years, several wireless wearable sensors have been developed to measure various parameters in neonates such as ECG, HR, RR, peripheral oxygen saturation and (skin) temperature.(8-14) Recently, a novel wireless and non-adhesive sensor belt (Bambi® belt, Bambi B.V., Eindhoven, the Netherlands) was developed for neonatal use that measures ECG and respiration based on the dEMG technique. A recent pilot study showed that measuring HR and RR with this belt in preterm infants is feasible and that the measured HR and RR trend was similar to ECG and CI.(submitted for publication, NICU AmsterdamUMC, 2021) However, before replacing the current techniques using adhesive wired electrodes with the non-adhesive sensor belt, a larger study is required to demonstrate the non-inferiority of this belt as an alternative cardiorespiratory monitor. In this study, we compare the monitoring performance of the Bambi® belt to ECG and CI and hypothesize that the performance of the belt is non-inferior to the current monitoring techniques.

METHODS

Study design

This multi-center paired non-inferiority study will be performed in the NICU of Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital of the Amsterdam University Medical Centre (AmsterdamUMC), both located in the Netherlands. Each patient will be simultaneously measured with the belt and ECG/CI (paired design). To compare the devices, a non-inferiority/equivalence framework will be used. Here, equivalence is defined as

the limit of agreement of the HR/RR between the belt and ECG/CI being within prespecified ~~clinically accepted~~ margins (see Table 1 for the margins). Non-inferiority is defined as the performance of clinical event detection and quality criteria not being worse than prespecified ~~clinically accepted~~ margins.

Study population

Preterm and term infants being routinely monitored with the standard cardiorespiratory monitor (Intellivue MP90, Philips Healthcare, Eindhoven, The Netherlands) are included in the study. To ensure a representative sample of the target population, infants in different age groups will be included. Infants with chest skin lesions, congenital anomalies, and other scenario's preventing belt placement, such as (effects of) surgery or wrap for therapeutic hypothermia, will be excluded.

Primary outcome

As HR-monitoring is clinically most relied upon and both ECG and dEMG provide the HR by measuring cardiac electrical activity, while CI and dEMG measure respiration with a different technique, the HR is considered the primary outcome.(3, 15) This will be studied with three criteria, which will be compared to the prespecified margins in Table 1. 1) Reliable monitoring performance through second-to-second HR measurement agreement in terms of differences in measured HR -between the belt and the ECG/-CI monitoring. 2) The detection of a composite cardiac event consisting of bradycardia (HR < 100 beats per minute for at least five seconds)(16) and tachycardia (HR > 180 beats per minute for at least ten seconds)(17) between the belt and the ECG measured with adhesive electrodes. The minimal duration of a bradycardia or tachycardia will prevent the inclusion of technical errors (short drops or increases in the HR) in our analysis and is lower for bradycardia compared to tachycardia as bradycardias are shorter events.(1) The thresholds are empirically chosen to detect all low and high HR-values. 3) Non-inferior quality (percentage of time with HR recordings without data loss).

Moreover, we will perform subgroup analyses to investigate whether the HR measurement performance is consistent under different clinical activities (e.g. kangaroo care, feeding) and in the different age groups.

Secondary outcomes

The secondary outcome is the measured RR. This will be studied using the following three criteria, which will be compared to the prespecified margins in Table 1:

1) Comparing the trend ~~(10-minute moving average)~~ in RR values provided by the belt and CI, based on the difference in the 10-minute moving averages. The RR-trend is studied as this ~~as the trend~~ is used in the clinical practice to detect for example increases in RR over time as a marker of clinical deterioration of a patient~~is the primary clinical usage of respiratory monitoring~~.(3) Since CI is widely used for neonatal respiratory monitoring, it is used as the reference technique.
2) Next to comparing the RR-trend, ~~Comparing~~ the ability to detect apnea and tachypnea is studied as the detection of these critical respiratory events based on RR is another purpose of the respiratory monitoring. Clinically relevant apneas are considered when indicated by a RR < 20 breaths per minute ~~(to capture all periods of low breathing frequency)~~ measured with CI for at least 10 seconds, associated with a desaturation (arterial oxygen saturation as measured by pulse oximetry (SpO_2) <80% for at least 10 seconds) and/or bradycardia (HR <100 beats per minute for at least five seconds) (objective apnea measurement).(16) A RR<20 breaths per minute is chosen for the apnea definition as we solely use the numerical RR-values, because despite the two different measurement techniques this endpoint is equal, and to capture all periods of low breathing frequency.

   Tachypnea is defined as a prolonged period of the averaged (moving average with a window size of 10 minutes) RR >60 breaths per minute and >100 breaths per minute (approximately two times the average normal RR).(18) To cover short and long periods of tachypnea, 3 different durations are studied (30 seconds, 60 seconds, and 10 minutes).

3) Calculating the percentage of time with reliable respiratory monitoring (without data loss and with an acceptable signal-to-noise ratio).

Data collection

The following basic characteristics and demographic information will be collected at the baseline of the study: gestational age, birth weight, gender, age and weight at day of measurement, relevant medical status (respiratory support, medication and underlying illness during measurement), chest circumference, nipple distance, skin type at study start by visual inspection (normal, dry, flaky, oily, moist, other).

Sample size calculation

A power calculation is performed for the primary outcome using data collected in a previous study.(submitted for publication, NICU AmsterdamUMC, 2021) Among the three criteria, criteria 1 needs the largest sample size and is used for our study. This resulted in 39 required infants to achieve 80% power with an overall 5% type I error with a Bonferroni correction (details in the Statistical Analysis Plan (SAP) in the online supplement).

In addition, an interim analysis will be performed as the power calculation was based on the previous study and recruitment of infants without being able to answer research questions is unethical.(19) This will be performed after including 1/3th of the infants for sample size adaption using the method of Mehta and Pocock.(20) If the conditional power falls within the pre-defined "promising zone", the sample size will be increased to an upper limit of 52 infants. Otherwise, the study will proceed with the original sample size.

To ensure that a representative sample of the age distribution of infants at a NICU, infants in different postmenstrual age groups will be recruited with the same proportions as in the target population (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks).

Study procedures

The Bambi® bBelt sSystem is a non CE-certified medical device, designed for wireless cardiorespiratory monitoring of (pre)term infants in a hospital environment. All included infants will be monitored with the belt in addition to standard ECG/CI for 24 hours to obtain representative clinical scenarios throughout the entire day. The measurement set-up is visualised in Figure 1 and consists of 1) dEMG measurement with the belt and 2) the extraction of patient monitor data.

In the belt, three dry electrodes are incorporated. When placing the belt at the height of the diaphragm, the outer two electrodes are in the nipple line and the middle electrode is in line with the sternum. The three ECG/CI electrodes are attached at the original location without hindering belt placement. The measured electrical signal of the diaphragm with the belt is wirelessly transmitted to the Receiver Module (REM) by the Sensor Module (SEM). The REM processes the dEMG signal to obtain the ECG and respiration signal (averaged diaphragmatic activity). An inbuilt algorithm provides the HR and RR out of the ECG and respiration signal respectively. This data is transported to a bedside computer. The data from the patient monitor (ECG, HR, RR, and SpO2) is extracted from the bedside monitor using an isolated cable and is also transported to the bedside computer.

The belt data from the REM and patient monitor are recorded and synchronised using a dedicated software package (Polybench, Applied Biosignals, Weener, Germany) on a personal bedside computer. Data is recorded at a sample rate of 1 to 500 Hz for rate and waveform data respectively. The bedside software also provides the possibility to make measurements annotations by nurses and researchers during data recording, such as re-positioning of the infant, nursing and kangaroo care.

During the study, daily routine care proceeds as usual. The location of the belt is regularly checked and if necessary repositioned (similar to the clinical practice). Notifications are visualised when contact between skin and the belt is lost (Leads off) or when there is no connection between the SEM and REM (Bluetooth Loss Error). In case of the first notification, the belt may be repositioned, while in case of Bluetooth loss the battery level of the SEM or blocking of this sensor (e.g. by an arm) are checked.

Preferably, the belt stays in place during the study. However, the belt can be removed during diagnostic imaging, patient handling, or in case of skin irritation at the belt location. The reason for removal will be annotated. If the belt is removed, the medical staff, parents and one of the dedicated researchers will decide together if the belt can be re-applied.

Recruitment

Parents of all eligible infants are approached for consent to obtain a sample as heterogeneous and representative as possible. Preferably, infants are included as soon as possible after birth. During the 24 hours, the study can be terminated if requested by parents or the treating physicians. In case of withdrawal of a subject, an extra subject will be included.

Safety

Being a medical device study, this study was classified as a moderate risk.(21) A specified monitor plan for the study is made based on risk-classification.

STATISTICAL ANALYSIS

A detailed SAP can be found in the online supplement. Unless otherwise specified, all hypothesis tests are two-sided with a significance level of 0.05. All statistical analyses will be

performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

The non-inferiority/equivalence margins based on expert opinions (survey send to neonatologists of different NICU's in the Netherlands) and literature (4, 22, 23) are described in Table 1. In the different subparagraphs we refer to this table.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***~~acceptance~~ Prespecified margins#***
LOA of second-to-second HR differences	± 8 bpm
LOA of RR trend differences	± 15 brpm
Sensitivity of brady-/tachycardia detection	90%°
PPV of brady-/tachycardia detection	90%°
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data with~~out~~ "Leads off" or "Bluetooth Loss Error" in the belt.

The ~~acceptance~~ prespecified margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

°Note: all missed bradycardias are checked for clinical relevance by two independent experts.

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac interference, all values for PPV for apnea/tachypnea are acceptable. Interpretations will be made

Summary and descriptive statistics

Categorical data will be summarized by numbers of counts and percentages. Continuous data will be summarized by mean, standard deviation if data is normal and median, interquartile range (IQR) if data is skewed. Minimum and maximum values will also be presented for continuous data when appropriate.

Statistical analysis of the primary outcome

Criteria 1: agreement in HR

To investigate the equivalence of HR measurement between the belt and ECG, we will fit a linear mixed model to the second-to-second HR difference between both. With this model, the 95% limits of agreement (Bland-Altman analysis) will be derived. The two-one-sided tests (TOST) with a multiplicity corrected alpha of 0.0167 and the prespecified margin (± 8bpm) will test equivalence between the two devices. In addition, based on a bivariate heteroscedastic model fitted to HR segments of a prespecified length, additional performance measures will be calculated as sensitivity analyses (details in SAP).

Criteria 2: cardiac event detection

For HR monitoring, we also consider the detection of bradycardia and tachycardia. We will estimate the sensitivity and the positive predictive value (PPV) of the belt using the patient monitor data as the ground truth and perform a non-inferiority test with an alpha of 0.0167. The non-inferiority margin for the sensitivity and PPV are listed in Table 1. In case of missed bradycardias, one independent expert per center will qualify the ~~severity~~ safety and clinical consequences ~~and acceptability~~ of each missing event by answering the same questions per figure containing the ~~with the use of the~~ discrepancy in HR and the ECG-signals measured with CI and the belt ~~and contextual information~~. ~~-~~ These figures will be blinded and thus it will be unknown which signal corresponds CI or the belt.

Criteria 3: signal quality

The quality of the investigational device will be quantified based on the percentage of time during the 24-hour period it produces any reading (percentage without data loss due to "Leads off" or "Bluetooth Loss Error") and the percentage in time it produces a good-quality-reading (percentage of robust data) for the HR and RR, respectively. For the HR non-robust data can be caused by bad connection (suboptimal Bluetooth or skin-electrode connection). These criteria are built-in in the belt algorithm and therefore this data is automatically labeled. Hypothesis testing will be used to establish the non-inferiority of this "uptime" percentage (percentage without data loss and percentage of robust data) of the belt.

For the RR, the uptime percentage is also categorized as a) data readings without data loss and b) robust data readings, ~~i.e. good-quality readings with acceptable signal-to-noise ratio's undisturbed by among others suboptimal connection and patient handling~~i.e. readings without unrealistic (e.g. negative) values. Signal quality is only analyzed for the belt. However, these results are compared to prespecified margins, ~~Again, the prespecified margins for the HR and RR are~~ described in Table 1. As the HR monitored with CI is accurate and nearly continuous, while the RR is less relied upon and may be unreliable, the prespecified margin for the RR is lower than for the HR.

Statistical analysis of secondary outcomes

Secondary analyses, based on the same statistical methods for the criteria of the primary outcome, include all secondary endpoints (apnea and tachypnea detection, RR trend analysis (see SAP)) and evaluation during different scenarios.

ETHICS AND DISSEMINATION

The Medical Ethical Committee of the MMC (W21.042) and the Central Committee for Human Research in the Netherlands (CCMO, CCMO21/0167/PP) approved the study protocol (Version

2, 19th of May 2021). Local feasibility at the AmsterdamUMC was approved by the Medical Ethical Committee of the AMC (2021_146**).** This study was registered in the Dutch Trial Register (https://www.trialregister.nl, NL9480). Regarding patient safety, no belt related events were observed in the pilot study and are therefore unexpected. Moreover, as every patient is monitored with ECG/CI and the belt, safety is guaranteed in case of missing belt data. The SAP will be used for the analyses. The results will be published in peer-reviewed journals and presented at future congresses.

The MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The duration of this study will be approximately seven months.

PATIENT AND PUBLIC INVOLVEMENT STATEMENT

Patients were included in this study after obtaining parental informed consent. The patients could not be involved in the design, recruitment, conduction and dissemination of results of this study. Neither could we ask the burden of the study. The outcome measures were developed by combining clinical and statistical knowledge to ensure a SAP that enables confirmation of non-inferiority of the belt compared to ECG/CI.

REFERENCES

1. Joshi R, van Straaten H, van de Mortel H, et al. Does the architectural layout of a NICU affect alarm pressure? A comparative clinical audit of a single-family room and an open bay area NICU using a retrospective study design. *BMJ Open* 2018;8(6):e022813. doi: 10.1136/bmjopen-2018-022813

2. Poets CF, Roberts RS, Schmidt B, Whyte RK, Asztalos E V, et al. Association Between Intermittent Hypoxemia or Bradycardia and Late Death or Disability in Extremely Preterm

Infants. *JAMA*. (2015) 314(6):595–603. https://doi.org/10.1001/jama.2015.8841~~Bouwstra S, Chen W, Feijs LMG, et al. Designing the parents-to-infant bonding experience in the NICU. *IARIA* 2012:7-13.~~

3.

3. Di Fiore JM. Neonatal cardiorespiratory monitoring techniques. *Semin Neonatol* 2004;9(3):195-203. doi: 10.1016/j.siny.2003.11.009

4. Kraaijenga JV, Hutten GJ, de Jongh FH, et al. Transcutaneous electromyography of the diaphragm: A cardio-respiratory monitor for preterm infants. *Pediatr Pulmonol* 2015;50(9):889-95. doi: 10.1002/ppul.23116

5. Csoma ZR, Meszes A, Abraham R, et al. Iatrogenic Skin Disorders and Related Factors in Newborn Infants. *Pediatr Dermatol* 2016;33(5):543-8. doi: 10.1111/pde.12960

6. Boundy EO, Dastjerdi R, Spiegelman D, et al. Kangaroo Mother Care and Neonatal Outcomes: A Meta-analysis. *Pediatrics* 2016;137(1). doi: 10.1542/peds.2015-2238

7. Conde-Agudelo A, Diaz-Rossello JL. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *Cochrane Database Syst Rev* 2016(8). doi: https://doi.org/10.1002/14651858.CD002771.pub4

8. Chung HU, Rwei AY, Hourlier-Fargette A, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. *Nat Med* 2020;26(3):418-29. doi: https://doi.org/10.1038/s41591-020-0792-9

9. Henry C, Shipley L, Ward C, et al. Accurate neonatal heart rate monitoring using a new wireless, cap mounted device. *Acta Paediatr* 2021;110(1):72-8. doi: https://doi.org/10.1111/apa.15303

10. Rimet Y, Brusquet Y, Ronayette D, et al. Surveillance of infants at risk of apparent life threatening events (ALTE) with the BBA bootee: a wearable multiparameter monitor. *Annu Int Conf IEEE Eng Med Biol Soc* 2007;2007:4997-5000. doi: 10.1109/IEMBS.2007.4353462

11. Chen H, Bao S, Lu C, et al. Design of an Integrated Wearable Multi-Sensor Platform Based on Flexible Materials for Neonatal Monitoring. *IEEE* 2020:23732-47. doi: 10.1109/ACCESS.2020.2970469

12. Chen W, Dols S, Bambang Oetomo S, et al. Monitoring body temperature of newborn infants at neonatal intensive care units using wearable sensors. *BodyNets '10* 2010. doi: https://doi.org/10.1145/2221924.2221960

13. Linti C, Horter H, Osterreicher P, et al. Sensory baby vest for the monitoring of infants. *IEEE* 2006. doi: 10.1109/BSN.2006.49

14. Pike H, Eilevstjonn J, Bjorland P, et al. Heart rate detection properties of dry-electrode ECG compared to conventional 3-lead gel-electrode ECG in newborns. *BMC Res Notes* 2021;14(1):166. doi: 10.1186/s13104-021-05576-x

15. van Leuteren RW, Hutten GJ, de Waal CG, et al. Processing transcutaneous electromyography measurements of respiratory muscles, a review of analysis techniques. *Journal of Electromyography and Kinesiology* 2019:176-86. doi: 10.1016/j.jelekin.2019.07.014

16. Lee H, Rusin CG, Lake DE, et al. A new algorithm for detecting central apnea in neonates. *Physiol Meas* 2012;33(1):1-17. doi: 10.1088/0967-3334/33/1/1

17. Li T, Matsushima M, Timpson W, et al. Epidemiology of patient monitoring alarms in the neonatal intensive care unit. *J Perinatol* 2018;38(8):1030-8. doi: 10.1038/s41372-018-0095-x

18. Reuter S, Moser C, Baack M. Respiratory distress in the newborn. *Pediatr Rev* 2014;35(10):417-28. doi: 10.1542/pir.35-10-417

19. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

20. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

21. Central Committee on Research Involving Human Subjects.

22. Anton O, Fernandez R, Rendon-Morales E, et al. Heart Rate Monitoring in Newborn Babies: A Systematic Review. *Neonatology* 2019;116(3):199-210. doi: 10.1159/000499675

23. Breteler MJM, KleinJan EJ, Dohmen DAJ, et al. Vital Signs Monitoring with Wearable Sensors in High-risk Surgical Patients: A Clinical Validation Study. *Anesthesiology* 2020;132(3):424-39. doi: 10.1097/ALN.0000000000003029

AUTHOR CONTRIBUTIONS

AS, HN, MV, RL, FJ, AK, JH conceptualized the study. ZZ and EH made the statistical analysis plan, which was reviewed by all authors. AS wrote the first version of this manuscript. All authors contributed to the final draft of the manuscript.

FUNDING STATEMENT

This work was supported by the Louise Vehmeijer foundation.

COMPETING INTERESTS STATEMENT

The authors declare that the research was conducted in the absence of any commercial ~~or financial~~ relationships that could be construed as a potential conflict of interest. However, the MMC and AmsterdamUMC do receive a prespecified amount of money per performed measurement. The results of this study protocol will contribute to obtaining a CE-mark for the Bambi® belt.

WHAT IS KNOWN ABOUT THE SUBJECT:

- Disadvantages of the cardiorespiratory monitoring technique in neonates are indirect measurements of respiration, usage of adhesive electrodes and hindering wires.
- With transcutaneous electromyography of the diaphragm, respiratory activity is measured directly by recording the activity of the main respiratory muscle.
- The Bambi® belt is a novel wireless and non-adhesive belt that enables cardiorespiratory monitoring by measuring diaphragm activity with dry electrodes.

WHAT THIS STUDY HOPES TO ADD:

- Demonstration of the non-inferiority of the Bambi® belt compared to the electrocardiogram and chest impedance for cardiorespiratory monitoring in preterm and term infants.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- When non-inferiority of the Bambi® belt compared to the current cardiorespiratory monitor is confirmed, the belt could be used as a wireless and skin-friendly alternative.

FIGURES

FIGURE 1 - The measurement set-up. The adhesive electrodes used for standard cardiorespiratory monitoring are attached at the original location, visualised by the three blue dots. The diaphragm activity measured with the Bambi® belt is wirelessly transmitted with the Sensor Module to the Receiver Module where the data is processed to obtain an electrocardiogram and respiration waveform (and heart rate and respiratory rate). This data and the data measured with the patient monitor are transported to a personal bedside computer with Polybench software to synchronise and record these signals.

MULTI-CENTER PAIRED NON-INFERIORITY STUDY OF THE CARDIORESPIRATORY MONITORING PERFORMANCE OF THE WIRELESS AND NON-ADHESIVE BAMBI® BELT MEASURING DIAPHRAGM ACTIVITY IN NEONATES: STUDY PROTOCOL

Anouk W.J. Scholten[1,2*†], Zhan Zhuozhao[3†], Hendrik J. Niemarkt[4], Marieke Vervoorn[4], Ruud W. van Leuteren[1,2], Frans H. de Jongh[1,5], Anton H. van Kaam[1,6], Edwin R. van den Heuvel[3], G. Jeroen Hutten[1,6]

[1]*Department of neonatology, Emma Children's Hospital, AmsterdamUMC, University of Amsterdam, Amsterdam, The Netherlands*

[2]*Amsterdam Reproduction & Development research institute, Amsterdam, The Netherlands*

3Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[4]*Department of neonatology, Màxima Medical Center, Veldhoven, The Netherlands*

[5]*Faculty of Science and Technology, University of Twente, Enschede, The Netherlands*

[6]*Department of neonatology, Emma Children's Hospital, AmsterdamUMC, Vrije Universiteit, Amsterdam, the Netherlands*

† These authors have contributed equally to this work and share first authorship

CORRESPONDING AUTHOR:

Anouk W.J. Scholten

Department of Neonatology

Emma Children's Hospital, AmsterdamUMC, location AMC

Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

Phone: +31 20 5663675

Email: a.w.j.scholten@amsterdamumc.nl

ABSTRACT

Introduction: Cardiorespiratory monitoring is used in the Neonatal Intensive Care Unit (NICU) to assess the clinical status of newborn infants and detect critical deteriorations in cardiorespiratory function. Currently, heart rate is monitored by electrocardiography (ECG) and respiration by chest impedance (CI). Disadvantages of current monitoring techniques are usage of wired adhesive electrodes which may damage the skin and hinder care. The Bambi® belt is a wireless and non-adhesive alternative that enables cardiorespiratory monitoring by measuring electrical activity of the diaphragm via transcutaneous electromyography (dEMG). A previous study showed feasibility of the Bambi belt and this study compares the belt performance to ECG and CI.

Methods and analysis: This multi-center non-inferiority paired study will be performed in the NICU of the Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital AmsterdamUMC in Amsterdam, the Netherlands. 39 infants in different postmenstrual age groups (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks) will be recruited. These infants will be monitored with the Bambi® belt in addition to standard ECG and CI for 24 h. The primary outcome is the heart rate (HR), studied with three criteria: 1) the limits of agreement of the HR measurements in terms of the second-to-second difference in the HR between the belt and standard ECG, 2) the detection of cardiac events consisting of bradycardia and tachycardia and 3) the quality of HR-monitoring. The secondary outcome is the respiratory rate (RR), studied with the criteria 1) agreement in RR trend monitoring, 2) apnea and tachypnea detection, and 3) reliable registrations.

Ethics and dissemination: This protocol was approved by the Medical Ethical Committee of the Màxima Medical Center and the Central Committee for Human Research (CCMO). The

MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The results will be presented at conferences and published in peer-reviewed journals.

Trial registration number: NL9480 (www.trialregister.nl)

INTRODUCTION

In the Neonatal Intensive Care Unit (NICU), cardiorespiratory monitoring is crucial to assess clinical condition and to timely detect and treat frequently occurring cardio-respiratory events to prevent morbidity and mortality.(1, 2) To date, this is performed by measuring the electrocardiogram (ECG) and chest impedance (CI) with three wired adhesive electrodes. CI measures variation in electrical impedance across the chest during respiration caused by changes in lung aeration and chest wall movement. These techniques provide continuous monitoring of heart rate (HR), respiratory rate (RR), and breathing pattern. However, as CI measures respiration indirectly, adequate detection of breathing cycles and apnea may not always be optimal.(3)

With transcutaneous electromyography of the diaphragm (dEMG) breathing effort can be recorded directly by measuring the electrical activity of this main respiratory muscle. To date, this technique also uses three adhesive electrodes and provides information on respiration and HR. Studies have shown its feasibility in the NICU-setting.(4)

The use of adhesive electrodes is restricted in infants with a postmenstrual age <26 weeks in fear of skin damage.(5) Moreover, electrode removal may cause discomfort. Furthermore, the wires attached to the electrodes restrict movements of the infant and may hinder parent-infant interaction, nursing and kangaroo care. Restrictions in kangaroo care may impact patient outcome as it has been associated with beneficial effects such as decreased mortality, decreased

risk of severe infection/sepsis and hypothermia, and increased likelihood of exclusive breast feeding.(6, 7) All things considered, it is important to find alternatives for using wired adhesive electrodes.

In the past years, several wireless wearable sensors have been developed to measure various parameters in neonates such as ECG, HR, RR, peripheral oxygen saturation and (skin) temperature.(8-14) Recently, a novel wireless and non-adhesive sensor belt (Bambi® belt, Bambi B.V., Eindhoven, the Netherlands) was developed for neonatal use that measures ECG and respiration based on the dEMG technique. A recent pilot study showed that measuring HR and RR with this belt in preterm infants is feasible and that the measured HR and RR trend was similar to ECG and CI.(submitted for publication, NICU AmsterdamUMC, 2021) However, before replacing the current techniques using adhesive wired electrodes with the non-adhesive sensor belt, a larger study is required to demonstrate the non-inferiority of this belt as an alternative cardiorespiratory monitor. In this study, we compare the monitoring performance of the Bambi® belt to ECG and CI and hypothesize that the performance of the belt is non-inferior to the current monitoring techniques.

METHODS

Study design

This multi-center paired non-inferiority study will be performed in the NICU of Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital of the Amsterdam University Medical Centre (AmsterdamUMC), both located in the Netherlands. Each patient will be simultaneously measured with the belt and ECG/CI (paired design). To compare the devices, a non-inferiority/equivalence framework will be used. Here, equivalence is defined as the limit of agreement of the HR/RR between the belt and ECG/CI being within prespecified

margins (see Table 1 for the margins). Non-inferiority is defined as the performance of clinical event detection and quality criteria not being worse than prespecified margins.

Study population

Preterm and term infants being routinely monitored with the standard cardiorespiratory monitor (Intellivue MP90, Philips Healthcare, Eindhoven, The Netherlands) are included in the study. To ensure a representative sample of the target population, infants in different age groups will be included. Infants with chest skin lesions, congenital anomalies, and other scenario's preventing belt placement, such as (effects of) surgery or wrap for therapeutic hypothermia, will be excluded.

Primary outcome

As HR-monitoring is clinically most relied upon and both ECG and dEMG provide the HR by measuring cardiac electrical activity, while CI and dEMG measure respiration with a different technique, the HR is considered the primary outcome.(3, 15) This will be studied with three criteria, which will be compared to the prespecified margins in Table 1. 1) Reliable monitoring performance through second-to-second HR measurement agreement in terms of differences in measured HR between the belt and the ECG/CI monitoring. 2) The detection of a composite cardiac event consisting of bradycardia (HR < 100 beats per minute for at least five seconds)(16) and tachycardia (HR > 180 beats per minute for at least ten seconds)(17) between the belt and the ECG measured with adhesive electrodes. The minimal duration of a bradycardia or tachycardia will prevent the inclusion of technical errors (short drops or increases in the HR) in our analysis and is lower for bradycardia compared to tachycardia as bradycardias are shorter events.(1) The thresholds are empirically chosen to detect all low and high HR-values. 3) Non-inferior quality (percentage of time with HR recordings without data loss).

Moreover, we will perform subgroup analyses to investigate whether the HR measurement performance is consistent under different clinical activities (e.g. kangaroo care, feeding) and in the different age groups.

Secondary outcomes

The secondary outcome is the measured RR. This will be studied using the following three criteria, which will be compared to the prespecified margins in Table 1:

1) Comparing the trend in RR values provided by the belt and CI, based on the difference in the 10-minute moving averages. The RR-trend is studied as this is used in the clinical practice to detect for example increases in RR over time as a marker of clinical deterioration of a patient.(3) Since CI is widely used for neonatal respiratory monitoring, it is used as the reference technique.
2) Next to comparing the RR-trend, the ability to detect apnea and tachypnea is studied as the detection of these critical respiratory events based on RR is another purpose of the respiratory monitoring. Clinically relevant apneas are considered when indicated by a RR < 20 breaths per minute measured with CI for at least 10 seconds, associated with a desaturation (arterial oxygen saturation as measured by pulse oximetry (SpO_2) <80% for at least 10 seconds) and/or bradycardia (HR <100 beats per minute for at least five seconds) (objective apnea measurement).(16) A RR<20 breaths per minute is chosen for the apnea definition as we solely use the numerical RR-values, because despite the two different measurement techniques this endpoint is equal, and to capture all periods of low breathing frequency.

   Tachypnea is defined as a prolonged period of the averaged (moving average with a window size of 10 minutes) RR >60 breaths per minute and >100 breaths per minute

(approximately two times the average normal RR).(18) To cover short and long periods of tachypnea, 3 different durations are studied (30 seconds, 60 seconds, and 10 minutes).

3) Calculating the percentage of time with reliable respiratory monitoring (without data loss and with an acceptable signal-to-noise ratio).

Data collection

The following basic characteristics and demographic information will be collected at the baseline of the study: gestational age, birth weight, gender, age and weight at day of measurement, relevant medical status (respiratory support, medication and underlying illness during measurement), chest circumference, nipple distance, skin type at study start by visual inspection (normal, dry, flaky, oily, moist, other).

Sample size calculation

A power calculation is performed for the primary outcome using data collected in a previous study.(submitted for publication, NICU AmsterdamUMC, 2021) Among the three criteria, criteria 1 needs the largest sample size and is used for our study. This resulted in 39 required infants to achieve 80% power with an overall 5% type I error with a Bonferroni correction (details in the Statistical Analysis Plan (SAP) in the online supplement).

In addition, an interim analysis will be performed as the power calculation was based on the previous study and recruitment of infants without being able to answer research questions is unethical.(19) This will be performed after including 1/3th of the infants for sample size adaption using the method of Mehta and Pocock.(20) If the conditional power falls within the pre-defined “promising zone”, the sample size will be increased to an upper limit of 52 infants. Otherwise, the study will proceed with the original sample size. To ensure that a representative sample of the age distribution of infants at a NICU, infants in different postmenstrual age groups

will be recruited with the same proportions as in the target population (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks).

Study procedures

The Bambi® belt system is a non CE-certified medical device, designed for wireless cardiorespiratory monitoring of (pre)term infants in a hospital environment. All included infants will be monitored with the belt in addition to standard ECG/CI for 24 hours to obtain representative clinical scenarios throughout the entire day. The measurement set-up is visualised in Figure 1 and consists of 1) dEMG measurement with the belt and 2) the extraction of patient monitor data.

In the belt, three dry electrodes are incorporated. When placing the belt at the height of the diaphragm, the outer two electrodes are in the nipple line and the middle electrode is in line with the sternum. The three ECG/CI electrodes are attached at the original location without hindering belt placement. The measured electrical signal of the diaphragm with the belt is wirelessly transmitted to the Receiver Module (REM) by the Sensor Module (SEM). The REM processes the dEMG signal to obtain the ECG and respiration signal (averaged diaphragmatic activity). An inbuilt algorithm provides the HR and RR out of the ECG and respiration signal respectively. This data is transported to a bedside computer. The data from the patient monitor (ECG, HR, RR, and SpO2) is extracted from the bedside monitor using an isolated cable and is also transported to the bedside computer.

The belt data from the REM and patient monitor are recorded and synchronised using a dedicated software package (Polybench, Applied Biosignals, Weener, Germany) on a personal bedside computer. Data is recorded at a sample rate of 1 to 500 Hz for rate and waveform data respectively. The bedside software also provides the possibility to make measurements

annotations by nurses and researchers during data recording, such as re-positioning of the infant, nursing and kangaroo care.

During the study, daily routine care proceeds as usual. The location of the belt is regularly checked and if necessary repositioned (similar to the clinical practice). Notifications are visualised when contact between skin and the belt is lost (Leads off) or when there is no connection between the SEM and REM (Bluetooth Loss Error). In case of the first notification, the belt may be repositioned, while in case of Bluetooth loss the battery level of the SEM or blocking of this sensor (e.g. by an arm) are checked.

Preferably, the belt stays in place during the study. However, the belt can be removed during diagnostic imaging, patient handling, or in case of skin irritation at the belt location. The reason for removal will be annotated. If the belt is removed, the medical staff, parents and one of the dedicated researchers will decide together if the belt can be re-applied.

Recruitment

Parents of all eligible infants are approached for consent to obtain a sample as heterogeneous and representative as possible. Preferably, infants are included as soon as possible after birth. During the 24 hours, the study can be terminated if requested by parents or the treating physicians. In case of withdrawal of a subject, an extra subject will be included.

Safety

Being a medical device study, this study was classified as a moderate risk.(21) A specified monitor plan for the study is made based on risk-classification.

STATISTICAL ANALYSIS

A detailed SAP can be found in the online supplement. Unless otherwise specified, all hypothesis tests are two-sided with a significance level of 0.05. All statistical analyses will be performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

The non-inferiority/equivalence margins based on expert opinions (survey send to neonatologists of different NICU's in the Netherlands) and literature (4, 22, 23) are described in Table 1. In the different subparagraphs we refer to this table.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***Prespecified margins[#]***
LOA of second-to-second HR differences	± 8 bpm
LOA of RR trend differences	± 15 brpm
Sensitivity of brady-/tachycardia detection	90%˚
PPV of brady-/tachycardia detection	90%˚
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data with "Leads off" or "Bluetooth Loss Error" in the belt.

[#] The prespecified margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

˚Note: all missed bradycardias are checked for clinical relevance by two independent experts.

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac interference, all values for PPV for apnea/tachypnea are acceptable. Interpretations will be made

Summary and descriptive statistics

Categorical data will be summarized by numbers of counts and percentages. Continuous data will be summarized by mean, standard deviation if data is normal and median, interquartile range (IQR) if data is skewed. Minimum and maximum values will also be presented for continuous data when appropriate.

Statistical analysis of the primary outcome

Criteria 1: agreement in HR

To investigate the equivalence of HR measurement between the belt and ECG, we will fit a linear mixed model to the second-to-second HR difference between both. With this model, the 95% limits of agreement (Bland-Altman analysis) will be derived. The two-one-sided tests (TOST) with a multiplicity corrected alpha of 0.0167 and the prespecified margin ($\pm$ 8bpm) will test equivalence between the two devices. In addition, based on a bivariate heteroscedastic model fitted to HR segments of a prespecified length, additional performance measures will be calculated as sensitivity analyses (details in SAP).

Criteria 2: cardiac event detection

For HR monitoring, we also consider the detection of bradycardia and tachycardia. We will estimate the sensitivity and the positive predictive value (PPV) of the belt using the patient monitor data as the ground truth and perform a non-inferiority test with an alpha of 0.0167. The non-inferiority margin for the sensitivity and PPV are listed in Table 1. In case of missed bradycardias, one independent expert per center will qualify the safety and clinical consequences of each missing event by answering the same questions per figure containing the discrepancy in HR and the ECG-signals measured with CI and the belt. These figures will be blinded and thus it will be unknown which signal corresponds CI or the belt.

Criteria 3: signal quality

The quality of the investigational device will be quantified based on the percentage of time during the 24-hour period it produces any reading (percentage without data loss due to "Leads off" or "Bluetooth Loss Error") and the percentage in time it produces a good-quality-reading (percentage of robust data) for the HR and RR, respectively. For the HR non-robust data can be caused by bad connection (suboptimal Bluetooth or skin-electrode connection). These criteria are built-in in the belt algorithm and therefore this data is automatically labeled. Hypothesis testing will be used to establish the non-inferiority of this "uptime" percentage (percentage without data loss and percentage of robust data) of the belt.

For the RR, the uptime percentage is also categorized as a) data readings without data loss and b) robust data readings, i.e. readings without unrealistic (e.g. negative) values. Signal quality is only analyzed for the belt. However, these results are compared to prespecified margins, described in Table 1. As the HR monitored with CI is accurate and nearly continuous, while the RR is less relied upon and may be unreliable, the prespecified margin for the RR is lower than for the HR.

Statistical analysis of secondary outcomes

Secondary analyses, based on the same statistical methods for the criteria of the primary outcome, include all secondary endpoints (apnea and tachypnea detection, RR trend analysis (see SAP)) and evaluation during different scenarios.

ETHICS AND DISSEMINATION

The Medical Ethical Committee of the MMC (W21.042) and the Central Committee for Human Research in the Netherlands (CCMO, CCMO21/0167/PP) approved the study protocol (Version 2, 19th of May 2021). Local feasibility at the AmsterdamUMC was approved by the Medical Ethical Committee of the AMC (2021_146**).** This study was registered in the Dutch Trial Register (https://www.trialregister.nl, NL9480). Regarding patient safety, no belt related events

were observed in the pilot study and are therefore unexpected. Moreover, as every patient is monitored with ECG/CI and the belt, safety is guaranteed in case of missing belt data. The SAP will be used for the analyses. The results will be published in peer-reviewed journals and presented at future congresses.

The MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The duration of this study will be approximately seven months.

PATIENT AND PUBLIC INVOLVEMENT STATEMENT

Patients were included in this study after obtaining parental informed consent. The patients could not be involved in the design, recruitment, conduction and dissemination of results of this study. Neither could we ask the burden of the study. The outcome measures were developed by combining clinical and statistical knowledge to ensure a SAP that enables confirmation of non-inferiority of the belt compared to ECG/CI.

REFERENCES

1. Joshi R, van Straaten H, van de Mortel H, et al. Does the architectural layout of a NICU affect alarm pressure? A comparative clinical audit of a single-family room and an open bay area NICU using a retrospective study design. *BMJ Open* 2018;8(6):e022813. doi: 10.1136/bmjopen-2018-022813

2. Poets CF, Roberts RS, Schmidt B, Whyte RK, Asztalos E V, et al. Association Between Intermittent Hypoxemia or Bradycardia and Late Death or Disability in Extremely Preterm Infants. *JAMA*. (2015) 314(6):595–603. https://doi.org/10.1001/jama.2015.88413.

3. Di Fiore JM. Neonatal cardiorespiratory monitoring techniques. *Semin Neonatol* 2004;9(3):195-203. doi: 10.1016/j.siny.2003.11.009

4. Kraaijenga JV, Hutten GJ, de Jongh FH, et al. Transcutaneous electromyography of the diaphragm: A cardio-respiratory monitor for preterm infants. *Pediatr Pulmonol* 2015;50(9):889-95. doi: 10.1002/ppul.23116

5. Csoma ZR, Meszes A, Abraham R, et al. Iatrogenic Skin Disorders and Related Factors in Newborn Infants. *Pediatr Dermatol* 2016;33(5):543-8. doi: 10.1111/pde.12960

6. Boundy EO, Dastjerdi R, Spiegelman D, et al. Kangaroo Mother Care and Neonatal Outcomes: A Meta-analysis. *Pediatrics* 2016;137(1). doi: 10.1542/peds.2015-2238

7. Conde-Agudelo A, Diaz-Rossello JL. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *Cochrane Database Syst Rev* 2016(8). doi: https://doi.org/10.1002/14651858.CD002771.pub4

8. Chung HU, Rwei AY, Hourlier-Fargette A, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. *Nat Med* 2020;26(3):418-29. doi: https://doi.org/10.1038/s41591-020-0792-9

9. Henry C, Shipley L, Ward C, et al. Accurate neonatal heart rate monitoring using a new wireless, cap mounted device. *Acta Paediatr* 2021;110(1):72-8. doi: https://doi.org/10.1111/apa.15303

10. Rimet Y, Brusquet Y, Ronayette D, et al. Surveillance of infants at risk of apparent life threatening events (ALTE) with the BBA bootee: a wearable multiparameter monitor. *Annu Int Conf IEEE Eng Med Biol Soc* 2007;2007:4997-5000. doi: 10.1109/IEMBS.2007.4353462

11. Chen H, Bao S, Lu C, et al. Design of an Integrated Wearable Multi-Sensor Platform Based on Flexible Materials for Neonatal Monitoring. *IEEE* 2020:23732-47. doi: 10.1109/ACCESS.2020.2970469

12. Chen W, Dols S, Bambang Oetomo S, et al. Monitoring body temperature of newborn infants at neonatal intensive care units using wearable sensors. *BodyNets '10* 2010. doi: https://doi.org/10.1145/2221924.2221960

13. Linti C, Horter H, Osterreicher P, et al. Sensory baby vest for the monitoring of infants. *IEEE* 2006. doi: 10.1109/BSN.2006.49

14. Pike H, Eilevstjonn J, Bjorland P, et al. Heart rate detection properties of dry-electrode ECG compared to conventional 3-lead gel-electrode ECG in newborns. *BMC Res Notes* 2021;14(1):166. doi: 10.1186/s13104-021-05576-x

15. van Leuteren RW, Hutten GJ, de Waal CG, et al. Processing transcutaneous electromyography measurements of respiratory muscles, a review of analysis techniques. *Journal of Electromyography and Kinesiology* 2019:176-86. doi: 10.1016/j.jelekin.2019.07.014

16. Lee H, Rusin CG, Lake DE, et al. A new algorithm for detecting central apnea in neonates. *Physiol Meas* 2012;33(1):1-17. doi: 10.1088/0967-3334/33/1/1

17. Li T, Matsushima M, Timpson W, et al. Epidemiology of patient monitoring alarms in the neonatal intensive care unit. *J Perinatol* 2018;38(8):1030-8. doi: 10.1038/s41372-018-0095-x

18. Reuter S, Moser C, Baack M. Respiratory distress in the newborn. *Pediatr Rev* 2014;35(10):417-28. doi: 10.1542/pir.35-10-417

19. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

20. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

21. Central Committee on Research Involving Human Subjects.

22. Anton O, Fernandez R, Rendon-Morales E, et al. Heart Rate Monitoring in Newborn Babies: A Systematic Review. *Neonatology* 2019;116(3):199-210. doi: 10.1159/000499675

23. Breteler MJM, KleinJan EJ, Dohmen DAJ, et al. Vital Signs Monitoring with Wearable Sensors in High-risk Surgical Patients: A Clinical Validation Study. *Anesthesiology* 2020;132(3):424-39. doi: 10.1097/ALN.0000000000003029

AUTHOR CONTRIBUTIONS

AS, HN, MV, RL, FJ, AK, JH conceptualized the study. ZZ and EH made the statistical analysis plan, which was reviewed by all authors. AS wrote the first version of this manuscript. All authors contributed to the final draft of the manuscript.

FUNDING STATEMENT

This work was supported by the Louise Vehmeijer foundation.

COMPETING INTERESTS STATEMENT

The authors declare that the research was conducted in the absence of any commercial relationships that could be construed as a potential conflict of interest. However, the MMC and AmsterdamUMC do receive a prespecified amount of money per performed

measurement. The results of this study protocol will contribute to obtaining a CE-mark for the Bambi® belt.

WHAT IS KNOWN ABOUT THE SUBJECT:

- Disadvantages of the cardiorespiratory monitoring technique in neonates are indirect measurements of respiration, usage of adhesive electrodes and hindering wires.
- With transcutaneous electromyography of the diaphragm, respiratory activity is measured directly by recording the activity of the main respiratory muscle.
- The Bambi® belt is a novel wireless and non-adhesive belt that enables cardiorespiratory monitoring by measuring diaphragm activity with dry electrodes.

WHAT THIS STUDY HOPES TO ADD:

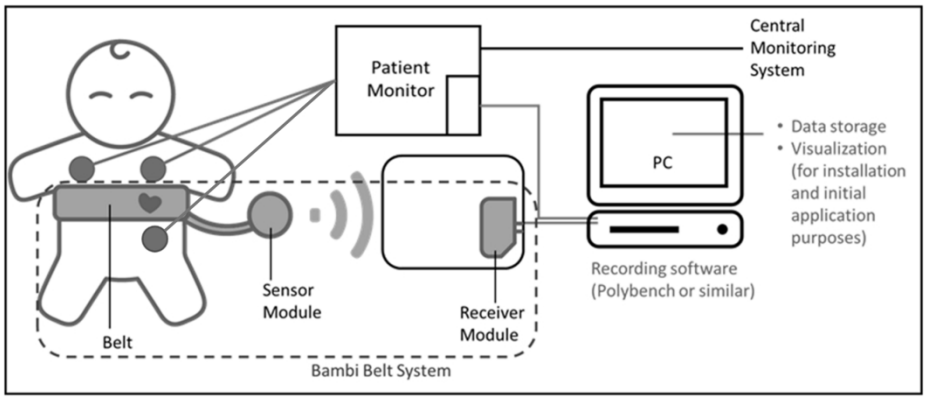
- Demonstration of the non-inferiority of the Bambi® belt compared to the electrocardiogram and chest impedance for cardiorespiratory monitoring in preterm and term infants.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- When non-inferiority of the Bambi® belt compared to the current cardiorespiratory monitor is confirmed, the belt could be used as a wireless and skin-friendly alternative.

FIGURES

FIGURE 1 - The measurement set-up. The adhesive electrodes used for standard cardiorespiratory monitoring are attached at the original location, visualised by the three blue dots. The diaphragm activity measured with the Bambi® belt is wirelessly transmitted with the Sensor Module to the Receiver Module where the data is processed to obtain an electrocardiogram and respiration waveform (and heart rate and respiratory rate). This data and the data measured with the patient monitor are transported to a personal bedside computer with Polybench software to synchronise and record these signals.



170x80mm (300 x 300 DPI)

ONLINE SUPPLEMENT – STATISTICAL ANALYSIS PLAN

PURPOSE OF THE STATISTICAL ANALYSIS PLAN

The purpose of this statistical analysis plan (SAP) is to document technical and detailed specification for the analysis of data collected in the Bambi belt monitoring performance study. The SAP has been written based on information contained in study protocol, dated 12th April 2021 before any data collection had taken place. It is prepared in compliance with the International Council on Harmonization (ICH) E9.

This SAP will be the guiding document for the analyses that will be conducted. Results of the analyses described in this SAP will be included in the Clinical Study Report (CSR). Any post hoc or unplanned analyses performed to provide results for inclusion in the CSR, but not identified in the prospective SAP will be identified in the given report. Additionally, the planned analyses of the primary aims will be included in future manuscripts. All the aims and research questions will be presented as an addendum as well.

OVERVIEW AND DESCRIPTION OF THE STUDY

Study design

The study is a multi-center, paired design, clinical monitoring device measurement comparison study. The *investigational device* under consideration is the Bambi® belt monitoring system (using dry electrodes). The current standard device of cardiorespiratory monitoring through adhesive electrodes is considered as the clinical reference standard and thereafter referred to as the *reference device/method*. The Bambi® belt monitoring system will be used on infants by trained nurses in the neonatal intensive care units (NICU's) for continuous 24 hours monitoring in addition to the routine monitoring with the reference device on the same patients. Infants admitted to NICU's of the the Emma Children's Hospital of the Amsterdam University Medical

Centre (Amsterdam UMC) or Maxima Medical Center (MMC) will be measured at the earliest suitable moment for clinical practice without interfering with infants' routine cycles.

Randomization and blinding

No randomization is required for the paired design since both monitoring devices will be used on the same patient at the same time. Blinding is also not possible since both the measurement protocol and algorithmic characteristic differ substantially.

Framework

The goal of this study is to establish the agreement between the investigational device and the reference device. Unlike the traditional difference-based tests, non-inferiority and equivalence techniques provide a better alternative for demonstrating the similarity between the two measurement methods. Thus, we have adopted the non-inferiority/equivalence trial framework for this primary objective of this study. This study considers three hypotheses (H_0 denotes the null hypothesis and H_A denotes the alternative hypothesis) for the first two primary outcomes:

1. Primary outcome, criterion 1: Heart rate measurement (second-by-second measurement)

H_0: The absolute difference between the investigational device and the reference device is larger than the prespecified equivalence margin.

H_A: The absolute difference between the investigational device and the reference device is within the prespecified equivalence margin.

2. Primary outcome, criterion 2: Brady-/tachy-cardia event detection

H_0: The composite cardiac event detection performances in terms of sensitivity and positive predictive value (PPV) based on the investigational device with respect to the reference device is less than the prespecified non-inferiority margin.

H_A: The composite cardiac event detection performances in terms of sensitivity and PPV based on the investigational device with respect to the reference device is greater or equal to the prespecified non-inferiority margin.

3. Primary outcome, criterion 3: Reliable reading (percentage of the time)

H_0: The percentage of the time the investigational device produces reliable readings is less than the prespecified non-inferiority margin.

H_A: The percentage of the time the investigational device produces reliable readings is greater or equal to the prespecified non-inferiority margin.

Statistical interim analysis and stopping guidance

One interim analysis for sample size adaptation will be performed. That is, we will start with a certain sample size commitment which will be increased at the interim analysis in case the results obtained are reasonably promising. The interim analysis will be conducted after the prospectively recruited participant's number reaches one-third of the planned sample size. Conditional power will be calculated for the analyses of the primary endpoints and compared to the boundary values of the conditional power for the promising zones (1, 2). In case the conditional power calculated at the interim analysis does fall inside the promising zone, the sample size will be increased to a predetermined limit. On the other hand, if the calculated conditional power is outside the promising zone, the study will proceed with the original sample size. Therefore, no early stopping rule is entailed in this study. Furthermore, a conventional

final analysis will be used without altering the level of type I error, since the promising zone is defined as a set that ensures the type I error to be preserved conservatively for the final analysis.

Study data

The following infant characteristics will be collected at baseline:

- Gestational age
- Postmenstrual age
- Gender
- Birth weight
- Weight at enrollment
- Ethnicity (derived from the electronic patient record or by asking the parents)
- Chest circumference
- Nipple distance
- Skin condition and abnormality

During the monitor study period, the following information will be measured:

- Clinical event
- SpO2: Arterial oxygen saturation as measured by pulse oximetry
- Medical status:
    - Ventilation support
    - Reports of medication and illness during the measurement
- Lead status: Indicates whether at least one lead was off
- Bluetooth link quality
- Activities
    - Kangaroo care
    - Nurse care

- Feeding
- Medical Procedure

- Belt status
  - Moved: the belt is being moved
  - Open: the belt is removed from the patient
- Patient position
  - Unknown
  - Lying prone
  - Lying supine
  - Lying on the left side
  - Lying on the right side

STATISTICAL ANALYSIS

Based on the collected information described above, the following total of variables will be derived:

- 10, 30, and 60 minutes moving average of the heart rate, and respiratory rate measured by both the investigational device and the reference device.
- Premature birth:
  - Premature (gestational age < 37 weeks)
  - Normal (gestational age >= 37 weeks)
- Desaturation: SpO2 < 80% for at least 10 consecutive seconds
- Heart rate status (investigational and reference device):
  - Normal
  - Tachycardia (heart rate > 180 for at least 10 consecutive seconds)
  - Bradycardia (heart rate < 100 for at least 5 consecutive seconds)

- Respiration status (investigational and reference device):
    - Apnea (according to standard clinical definitions)
    - Tachypnea (respiratory rate >60 and >100 for 30 seconds, 1 minute, and 10 consecutive minutes in stationary signal)
- Measurement quality:
    - No anomalies
    - Poor data link: Bluetooth link is poor but data is still received
    - Unreliable data: One or more lead off, or no Bluetooth connection (Bluetooth Loss Error, BLE)

Summary and descriptive statistics

Categorical data will be summarized by numbers and percentages. Continuous data will be summarized by mean, standard deviation if data are normal and median, interquartile range (IQR) if data are skewed. Minimum and maximum values will also be presented for continuous data. Tests of statistical significance will not be undertaken for baseline characteristics; rather the clinical importance of any imbalance will be noted.

A CONSORT flow diagram (example in Figure 1) will be used to summarize the number of infants who were:

- Assessed for eligibility at the screening
    - Eligible at screening
    - Ineligible at screening (with reasons)
- Eligible and enrolled
- Eligible but not enrolled
- Enrolled but did not receive any / sufficient measurements

- Discontinued

- Included in the analysis
- Excluded from the analysis

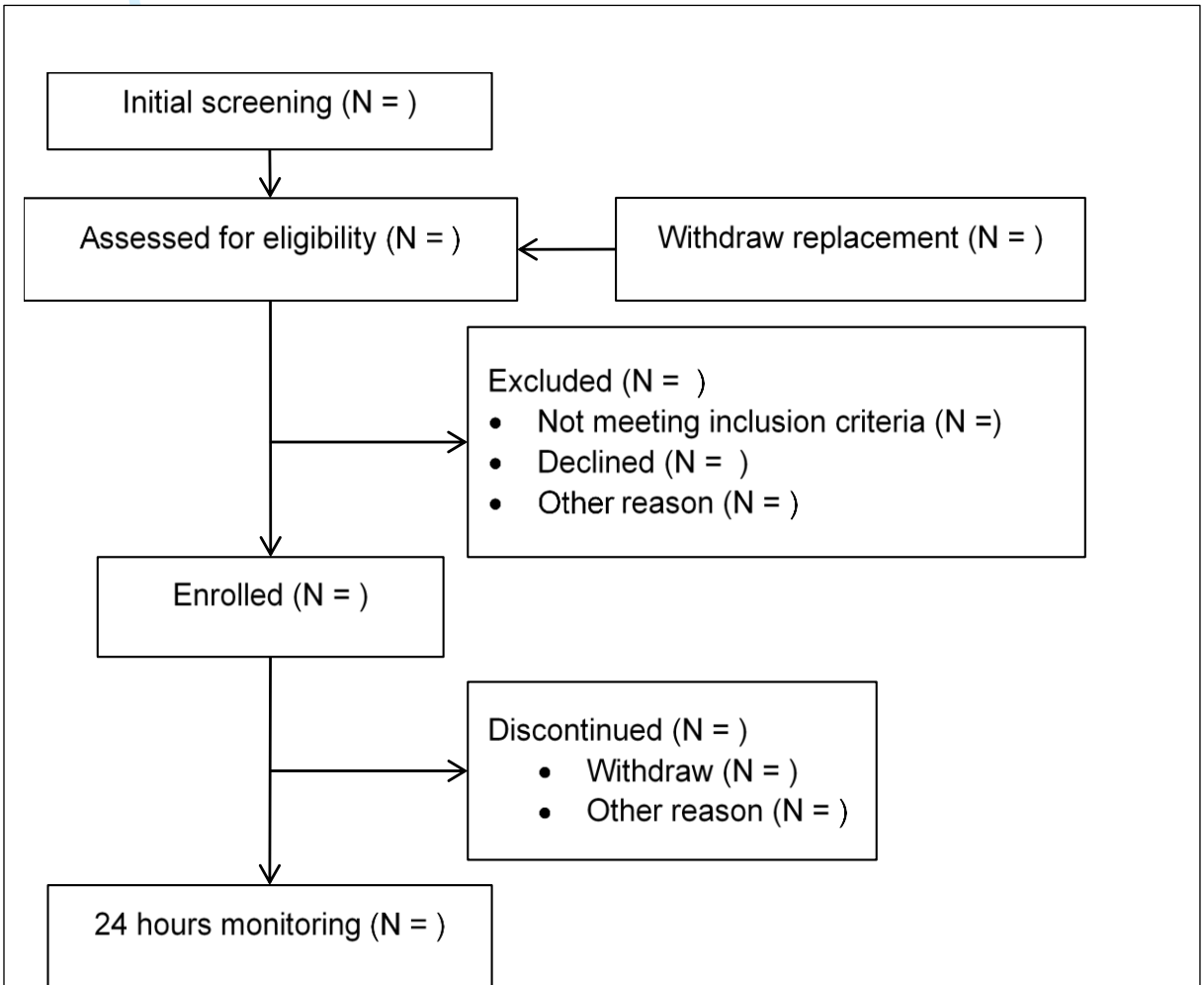


FIGURE 1 – Example of a CONSORT flow diagram

Analysis methods

Primary outcome, criterion 1: Heart rate measurement

To investigate and verify the equivalence of heart rate measurement between the investigational device and the reference device, we will fit a linear mixed model to the second-to-second heart rate difference between the two devices. Based on the estimates of the model, we will derive the 95% limits of agreement (3) as our main performance measure, known as the Bland-Alman analysis. The endpoints of the Bland-Altman 95% limits of agreement are the 2.5th percentile and 97.5th percentile for the distribution of the difference between paired measurements. We will calculate the $(1-\alpha/2)100\%$ confidence intervals of the percentiles according to Shieh (4), and conduct the two-one-sided t-tests (TOST) procedure with the prespecified equivalence margins (Table 1).

In addition, we will calculate the following performance measures to supplement the main analysis as sensitivity analyses to assess the agreement between the two devices from different aspects:

- The concordance correlation coefficient (5) and its variants
- Probability of Agreement (6) and Total Deviation Index (7)
- Coefficient of individual agreement (8)

These performance measures will be based on a bivariate heteroscedastic linear mixed-effects model fitted to each segment of the readings of a prespecified length from both devices. We will assume that measurements made with the two devices at the same time are correlated. Therefore, investigating the correlation between the two devices leads to the quantification of the degrees of agreement between them. Furthermore, we will consider the temporal correlations between measurements obtained with the same devices and the variabilities between different infants. Besides, we will start with a heteroscedastic model which does not assume equal variances for the two devices (namely, the measurement errors are not assumed

to be equal) and investigate the homogeneity of the measurement variabilities between the two devices. Baseline characteristics of the infants and records of activities (listed in the study data section) will be used as covariates in the model to partly explain the variabilities between the infants. We will use the stepwise model selection procedure based on the Bayesian Information Criteria (BIC) goodness-of-fit criteria.

Primary outcome, criterion 2: Brady-/tachycardia event detection

For brady-/tachycardia, the clinical event periods will be identified based on prespecified margins. We will investigate the non-inferiority of sensitivity and positive predictive values (PPV) of the event detected by the investigation device assuming that the reference device is the predicate device and compare both values to the prespecified non-inferiority margins (Table 1). For the calculation of the sensitivity, when the event period identified based on the investigational device overlaps with the event period identified by the reference device, it will be counted as a true positive case. This is to prevent the repeated signaling of events from the investigational device during a positive period identified by the reference device to inflate the number of true positives. The same applies to the reference device when it comes to the calculation of the PPV. That is, during an event period identified by the investigational device, multiple event periods identified by the reference device will only be counted as one true-positive case. Note that the true negative is ill-defined and will not be reported. Since true negatives are used in the calculation of specificity, specificity will not be reported either.

Primary outcome, criterion 3: Safety and Quality

Safety: The investigation of safety and tolerability is a multidimensional problem. Although we don't anticipate any specific adverse effects for the investigational device, new and unforeseeable effects are always possible. This background underlies the statistical difficulties

associated with the analytical evaluation of the safety and tolerability of the device. We will address the safety and tolerability implications by applying descriptive statistical methods to the data, supplemented by calculation of confidence intervals whenever this aids interpretation and make use of graphical presentations in which patterns of adverse events are displayed.

Quality: The quality of the investigational device will be quantified in terms of the point estimate and 95% confidence intervals based on the estimated percentages in time during the 24-hour period it produces reliable readings for heart rate and respiratory rate, respectively. Reliable readings are defined in the study protocol. The uptime percentages are the percentage of data loss and the percentage of robust data readings. For each outcome, hypothesis testing will be used to establish the non-inferiority of the uptime percentages of the investigational device considering a non-inferiority margin specified in Table 1. The uptime percentages will be estimated based on a Generalized Estimation Equations (GEE) model.

Missing data

To get an idea about the complexity of the missing data problem in the data and information about the location of the missing values, the missing data pattern will be evaluated and reported. We expect missing data in the primary outcomes measured by the investigational device to be the results of external causes such as the movement of the belt, signal losses, poor Bluetooth link qualities and so on. Therefore, it will be reasonable to assume that data are missing completely at random (MCAR). Formally, we will investigate the validity of such an assumption using Little's MCAR test. Furthermore, the availability of the data from the reference device (since it depends on a separate measurement system) provides us the opportunity to investigate whether the missingness is related to the underlying measurand. That is, whether the missing data mechanism is missing not at random (MNAR). This is rarely

possible in other types of studies. Nevertheless, considering the pair of bivariate measurements from the investigational and the reference device, we will investigate the assumption using the covariate-dependent missing (CDM) test proposed in Li (9). Note that CDM is usually considered as missing at random (MAR), we here simply exploit the advantage of the data from the reference device to test the dependencies between the missingness and the underlying measurand. Furthermore, we will use the CDM test on other covariates (excluding the reference device data) as well to test if the missingness is MAR.

In the case of MAR (i.e., CDM without measurements from reference device), list wise deletion can still be unbiased and will be used if the percentage of missingness is less than 5%. Otherwise, multiple imputations (MI) will be considered. We will not use the measurements from the reference device for the MI to avoid biasing the results towards the equivalence of the two devices. On the other hand, if the missingness is related to the measurand after taking into account all covariates, this indicates a potential problem of the measurement device, and a separate analysis will be carried out to investigate the associations between the missingness and the measurand.

For multiple imputations, we will use the fully conditional specification method. Unrealistic values (e.g., negative values for strictly positive variable) will be checked and corrected (e.g., using truncations). The imputation will be repeated at least 5 times and Rubin's rule will be used to combine estimates and standard errors from the imputed data.

Secondary analyses

If the sample size permits, we will perform subset analyses to explore the performances of the investigational device under different scenarios.

Subset analyses: primary endpoints

For each of the primary endpoints, we will consider additional exploratory analyses on the following subsets:

- During periods of a clinical event (e.g., apnea, bradycardia)
- During activities (e.g., Kangaroo care, feeding)
- During periods where the reference device's readings are stable
- Gestational age (e.g., preterm birth)
- Respiratory support (e.g., mechanical ventilation)

For these subsets, we will use the same model as the primary outcome to investigate the performances of the investigational device under various scenarios/activities of the infants. In case the subset does not contain enough data to fit the same model as the primary one, we will resort to a simpler model for case-by-case analyses.

Respiratory rate analysis

It is known the reference device does not provide point-by-point accurate measurement resulting in large variabilities (measurement errors) in the measured respiratory rates. The intended clinical use of the readings in the NICU thus consists of two different aspects:

1. The trend of the respiratory rates over time;
2. Signaling of potentially respiratory related clinical events (i.e. apnea related desaturation and/or bradycardia, and potentially disease related tachypnea);

For the first usage, we will apply the same analysis method as the one used for heart rate on the moving average of the respiratory rate. We will primarily focus on the 10 minutes moving average for the respiratory rate. Analysis of the 1 minute and 5 minutes moving averages will

be used as a sensitivity analysis to establish the robustness of the conclusions made for the 10 minutes moving average.

For apnea and tachypnea, respectively, the clinical event periods will be identified based on clinical definitions and the same methods as the brady-/tachycardia event detection will be used to compare the sensitivity and PPV to the prespecified non-inferiority limits (Table 1). However, it should be noted that since the reference device is known to have an unsatisfactory performance of apnea/tachypnea detection, cautions are needed to interpret the sensitivity and PPV as if the reference device is the truth.

Statistical software

All statistical analyses will be performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

Non-inferiority/equivalence criteria

In Table 1 the non-inferiority/equivalence criteria for the primary and secondary outcomes are visualized.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***Prespecified margins#***
LOA of second-to-second HR differences	±8 bpm
LOA of RR trend differences	±15 brpm
Sensitivity of brady-/tachycardia detection	90%˚
PPV of brady-/tachycardia detection	90%˚
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data with "Leads off" or "Bluetooth Loss Error" in the belt.

The prespecified margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

˚Note: all missed bradycardias are checked for clinical relevance by two independent experts.

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac

Sample size

Based on preliminary analysis of data collected in a feasibility study on a total of 13 infants with measurements from both the investigational device and the reference device, we were able to obtain preliminary information with regards to the characteristics of the primary endpoints

upon which we have formulated our sample size calculation.(submitted for publication, NICU AmsterdamUMC, 2021)

A detailed specification of the sample size calculation can be found in the sections below. In summary, for the monitor performance, 39 infants are needed to achieve 80% power with a 5% overall type I error with a Bonferroni correction for multiplicity. It is worth noting that no dropout was assumed during the sample size calculation. This is because we plan to include an extra infant in case of withdrawal of an infant to fulfil the required sample size. Infants who withdraw from the study will be followed up by one of the investigators and responsible medical staff to obtain detailed reasons behind the withdraw. Dropout rate for the monitor performance study is expected to be low, between 0-5%.

While the preplanned sample size is 39 infants, we will include an adaptive sample size re-estimation procedure as per the "promising zone" methodology of Mehta and Pocock (2) using the data from the first 1/3 infants. This procedure involves the evaluation of conditional power in the interim analysis, and if it were to fall in the prespecified "promising zone", the sample size will be increased, subject to a predetermined upper limit (52 infants) to increase the conditional power to 80%. The boundary of the conditional power for the "promising zone" is 0.36 and 0.8.

Monitoring study: Primary endpoints: Heart rate

For the sample size calculation, we will assume the measured heart rate difference D_{ij} between the investigational device and the reference device at time point j $(j = 1, \dots, m)$ on infant i $(i = 1, \dots, n)$ can be modelled as:

$$D_{ij} = d + a_i + e_{ij}$$

where d is the overall difference, a_i is a random effect with $a_i \sim N(0, \sigma_a^2)$, and $e_{ij} \sim N(0, \sigma_e^2)$ is the random error independent of a_i. Though, we considered a bivariate mixed-effects model for our analysis, the variance component model for the difference can be derived from the bivariate mixed-effects model, therefore we will use this variance component model for the sample size calculation. The variance of the difference will be estimated from the aforementioned model via $\hat{\sigma}_d^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2$. Here $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ is the estimator of the between-subject variability σ_a^2 and residual variability σ_e^2, respectively. The 95% limit of agreement (LOA) can be estimated as LOA = $\hat{d} \pm 1.96\,\hat{\sigma}_d$ with $\hat{d}$ and $\hat{\sigma}_d$ denotes the estimator of d and σ_d, respectively. The variance of the LOA estimator is $\text{var}\left(\hat{d} \pm 1.96\,\hat{\sigma}_d\right) = \text{var}\left(\hat{d}\right) + 1.96^2\text{var}(\hat{\sigma}_d)$ ($\hat{d}$ and $\hat{\sigma}_d$ is asymptotically independent). Since for $\hat{\sigma}_d^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2$, we have $\text{var}(\hat{\sigma}_d^2) = \text{var}(\hat{\sigma}_a^2) + \text{var}(\hat{\sigma}_e^2) + \text{cov}(\hat{\sigma}_a^2, \hat{\sigma}_e^2)$. Furthermore, each term on the right-hand side (assuming m is large) is given by:

$$var\,(\hat{\sigma}_a^2) = \frac{2}{m^2}\left[\frac{(m\sigma_a^2 + \sigma_e^2)^2}{n-1} + \frac{\sigma_e^4}{n(m-1)}\right] \approx \frac{2\sigma_a^4}{n-1}, var\,(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{n(m-1)+2} \approx 0,$$

$$cov(\hat{\sigma}_a^2, \hat{\sigma}_e^2) = -\frac{2\sigma_e^4}{nm(m-1)} \approx 0;$$

This leads to $\text{var}\left(\hat{\sigma}_d{}^2\right) \approx 2\sigma_a{}^2/(n-1)$. Therefore, by the delta method, we have $\text{var}(\hat{\sigma}_d) = \frac{1}{4\sigma_d{}^2} var(\hat{\sigma}_d{}^2) = \frac{\sigma_a{}^4}{2(n-1)\sigma_d{}^2}$. According to Lu et al. (10), the power for the TOST is given by:

$$1 - \beta = 1 - T_{n-1}\left(t_{1-\frac{\alpha}{2}}, \frac{\delta - d - 1.96\sigma_d}{se_{LOA}}\right) - T_{n-1}\left(t_{1-\frac{\alpha}{2}}, \frac{\delta + d - 1.96\sigma_d}{se_{LOA}}\right)$$

where α, β denotes type I and type II error respectively, δ is the predefined limit, $se_{LOA} \approx \sqrt{\frac{{\sigma_d}^2}{n} + \frac{1.96^2{\sigma_a}^4}{2(n-1){\sigma_d}^2}}$ is the standard error of the LOA estimate calculated according to the variance component model, and $T_{n-1}(\cdot, \tau)$ denotes the cumulative distribution function of a non-central Student's t-distribution with n-1 degrees of freedom, and non-centrality parameter τ.

For a 5% overall type I error rate, with a multiplicity correction factor of 3, and 80% power, the minimum sample size required is calculated at $n = 39$, for $d = -0.5$, $\sigma_a = 0.3$, and $\sigma_d = 3$.

Primary endpoints: Brady-/tachycardia event detection

Suppose the total number of true events is M and are 100% detected by the reference device. Assuming the true sensitivity is 95% for the investigational device, then a non-inferiority test using Z-test with normal approximation to the binomial distribution leads to a required M of 271 for a power of 80% and $\alpha = 0.05/3 \approx 0.01667$ assuming the detection between each event (conditioning on the event itself) is independent. Considering the incidence of bradycardia to be 1 event per hour per infant according to the preliminary analysis of data from the feasibility study, at least 12 infants are needed to satisfy the required M (assuming each infant is measured for 24 hours long). The calculation is the same for PPV if we assume the investigational device is the truth. Assuming an incidence rate of tachycardia of 1.5 per hour per infant according to the preliminary analysis of data from the feasibility study, the required sample size is 8. Note that in the aforementioned calculation, we assume that the event-detection performance of the investigational device is homogeneous (or independent) among infants. A

sensitivity/robustness investigation regarding the sample size for infant-specific heterogeneous performances was performed, with results from which we can see that with $n = 39$, we have more than 90% power to detect a heterogeneous performance scenario where 15% of the population would have sensitivities between 80% - 90% and less than 5% of the population have sensitivities less than 80%.

Primary endpoints: Safety and quality

Based on preliminary analysis of data collected in the feasibility study, we will assume that the overall probability of producing an erroneous reading at any time p_e is 2% and is constant across all participants. We will consider a non-inferiority test using normal approximation and a Z-test with the null hypothesis of $H_0: p_e > 0.05$ and the alternative hypothesis of $H_A: p_e \leq 0.05$. The required number of observations for a given type I error of 1.667% ($\approx 5\%/3$) to achieve 80% power is 376. Here the sample size 376 refers to 376 independent observations. Considering the large numbers of repeated measurements (more than 376) within each participant, we will have sufficient power for this non-inferiority test even with 1 participant. However, the assumption of independence can be too strong in the setting of our study. Therefore, if we would assume an AR(1)-type dependency with correlation parameter $\rho = 0.8$ between two measurements within a participant, the variance inflation factor (VIF) for the asymptotic variance of the GEE estimator $\hat{p}_e$ according to Pan is approximately (with the number of repeats $m = 376$):

$$1 + \frac{2\rho}{1-\rho} = 9$$

in the case of identity working correlation matrix when the true correlation has an AR(1) structure. To achieve the same power as the independent case calculated before, we need

$$var(\hat{p}_e) \coloneqq VIF\,\frac{p_e(1-p_e)}{nm} = \frac{p_e(1-p_e)}{m}$$

Thus, we can conclude that at least $n = \text{VIF} = 9$ participants will be needed to provide enough power for the non-inferiority test based on the GEE estimator using the identity working correlation matrix using the inverse proportionality between the required sample size and the variance of the estimator used in the Z-test. The same calculation can be carried out for the robust data percentages. It can be seen that only the number of repeats m will differ when the probabilities and the non-inferiority margins change while the VIF remains the same for the same value of the correlation parameter ρ. Among all settings, the largest m needed will be 718 when we assume the probability of producing robust data for respiratory rate is 75% with the corresponding non-inferiority margin equals to 70%. This number of repeats is still fully covered by the high-frequency measurements found in the study.

Protocol deviations and analysis sets

Definition of protocol deviations

Protocol deviations (PD) occurring during the study will be determined for all enrolled infants, mainly from the clinical database by either clinical and/or medical review processes.

The mapping of the protocol deviations from the clinical database to analysis will be performed as per Table 2:

Table 2 – The influence of protocol deviations on the statistical analysis plan (SAP)

Database label	*SAP*
Minor	Not required
Major	Important
Critical	Important
Clinical (a subset of Critical)	Important

Important protocol deviations are protocol deviations that might significantly affect the completeness, accuracy, and/or reliability of the study data or that might significantly affect a subject's rights, safety, or well-being.

Important protocol deviations may also be recorded as "Major" protocol deviations in the database, but will be presented only as important in the analysis output.

Important protocol deviations include:

- Infants that are included in the study despite not satisfying the eligibility criteria;
- Infants that develop exclusion criteria while on the study but not withdrawn;
- Infants being measured with operational human errors;
- Deviation from Good Clinical Practice (ICE E6)

Clinically Important protocol deviations are the protocol deviations marked as important in Table 2, which lead to the exclusion of a subject from the analysis set.

The following deviations will be identified and confirmed before the partial database lock for the final analysis.

- Important protocol deviations including
    - Deviations from the inclusion and exclusion criteria
    - Deviations post inclusion

Protocol deviations may be identified by the data managers, clinical and medical staff either by programmed validation checks or data listings/reports or manual verification of data sources. Some important/major protocol deviation criteria may be identified in the clinical database via biostatistical programs. Every important protocol deviation will be documented in the database whether identified through sites monitoring, medical review or programming.

REFERENCES

1. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

2. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

3. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8(2):135-60. doi: 10.1177/096228029900800204

4. Shieh G. The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. *BMC Med Res Methodol* 2018;18(1):45. doi: https://doi.org/10.1186/s12874-018-0505-y

5. Lin L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989:255-68.

6. Stevens NT, Steiner SH, MacKay RJ. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Stat Methods Med Res* 2017;26(6):2487-504. doi: 10.1177/0962280215601133

7. Lin L. I.-K. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 2000:255-70.

8. Haber M, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Stat Methods Med Res* 2008;17(2):151-69.

9. Li C. Little's Test of Missing Completely at Random. *The Stata Journal* 2013. p. 795-809. doi: 10.1177/1536867X1301300407

10. Lu MJ, Zhong WH, Liu YX, et al. Sample Size for Assessing Agreement between Two Methods of Measurement by Bland-Altman Method. *Int J Biostat* 2016;12(2). doi: 10.1515/ijb-2015-0039

BMJ Paediatrics Open

Multi-center paired non-inferiority study of the cardiorespiratory monitoring performance of the wireless and non-adhesive Bambi® belt measuring diaphragm activity in neonates: Study Protocol

Journal:	*BMJ Paediatrics Open*
Manuscript ID	bmjpo-2022-001430.R2
Article Type:	Protocol
Date Submitted by the Author:	25-Apr-2022
Complete List of Authors:	Scholten, Anouk; Amsterdam UMC Locatie AMC, department of Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute Zhuozhao, Zhan; Eindhoven University of Technology, department of Mathematics and Computer Science Niemarkt, Hendrik; Máxima Medisch Centrum, department of Neonatology Vervoorn, Marieke; Máxima Medisch Centrum, department of Neonatology van Leuteren, Ruud; Amsterdam UMC Locatie AMC, department of Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute de Jongh, Frans; Amsterdam UMC Locatie AMC, department of Neonatology; University of Twente, Faculty of Science and Technology van Kaam, Anton; Amsterdam UMC Locatie AMC, department of Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute van den Heuvel, Edwin; Eindhoven University of Technology, department of Mathematics and Computer Science Hutten, G. Jeroen; Amsterdam UMC Locatie AMC, department of Neonatology; Amsterdam UMC Locatie AMC, Amsterdam Reproduction & Development research institute
Keywords:	Neonatology, Technology

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

MULTI-CENTER PAIRED NON-INFERIORITY STUDY OF THE CARDIORESPIRATORY MONITORING PERFORMANCE OF THE WIRELESS AND NON-ADHESIVE BAMBI® BELT MEASURING DIAPHRAGM ACTIVITY IN NEONATES: STUDY PROTOCOL

Anouk W.J. Scholten[1,2*†], Zhan Zhuozhao[3†], Hendrik J. Niemarkt[4], Marieke Vervoorn[4], Ruud W. van Leuteren[1,2], Frans H. de Jongh[1,5], Anton H. van Kaam[1,2], Edwin R. van den Heuvel[3], G. Jeroen Hutten[1,2]

[1]*Amsterdam UMC location University of Amsterdam, department of Neonatology, Meibergdreef 9, Amsterdam, The Netherlands*

[2]*Amsterdam Reproduction & Development research institute, Amsterdam, The Netherlands*

[3]*Eindhoven University of Technology, department of Mathematics and Computer Science, Eindhoven, The Netherlands*

[4]*Màxima Medical Center, department of neonatology, Veldhoven, The Netherlands*

[5]*University of Twente, Faculty of Science and Technology, Enschede, The Netherlands*

† These authors have contributed equally to this work and share first authorship

CORRESPONDING AUTHOR:

Anouk W.J. Scholten

Department of Neonatology

Emma Children's Hospital, AmsterdamUMC, location AMC

Meibergdreef 9, 1105AZ Amsterdam, The Netherlands

Phone: +31 20 5663675

Email: a.w.j.scholten@amsterdamumc.nl

WORD COUNT: 2777

WHAT IS KNOWN ABOUT THE SUBJECT:

- Disadvantages of the cardiorespiratory monitoring technique in neonates are indirect measurements of respiration, usage of adhesive electrodes and hindering wires.
- With transcutaneous electromyography of the diaphragm, respiratory activity is measured directly by recording the activity of the main respiratory muscle.
- The Bambi® belt is a novel wireless and non-adhesive belt that enables cardiorespiratory monitoring by measuring diaphragm activity with dry electrodes.

WHAT THIS STUDY HOPES TO ADD:

- Demonstration of the non-inferiority of the Bambi® belt compared to the electrocardiogram and chest impedance for cardiorespiratory monitoring in preterm and term infants.

ABSTRACT

Introduction: Cardiorespiratory monitoring is used in the Neonatal Intensive Care Unit (NICU) to assess the clinical status of newborn infants and detect critical deteriorations in cardiorespiratory function. Currently, heart rate is monitored by electrocardiography (ECG) and respiration by chest impedance (CI). Disadvantages of current monitoring techniques are usage of wired adhesive electrodes which may damage the skin and hinder care. The Bambi® belt is a wireless and non-adhesive alternative that enables cardiorespiratory monitoring by measuring electrical activity of the diaphragm via transcutaneous electromyography (dEMG). A previous study showed feasibility of the Bambi belt and this study compares the belt performance to ECG and CI.

Methods and analysis: This multi-center non-inferiority paired study will be performed in the NICU of the Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital AmsterdamUMC in Amsterdam, the Netherlands. 39 infants in different postmenstrual age groups (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks) will be recruited. These infants will be monitored with the Bambi® belt in addition to standard ECG and CI for 24 h. The primary outcome is the heart rate (HR), studied with three criteria: 1) the limits of agreement of the HR measurements in terms of the second-to-second difference in the HR between the belt and standard ECG, 2) the detection of cardiac events consisting of bradycardia and tachycardia and 3) the quality of HR-monitoring. The secondary outcome is the respiratory rate (RR), studied with the criteria 1) agreement in RR trend monitoring, 2) apnea and tachypnea detection, and 3) reliable registrations.

Ethics and dissemination: This protocol was approved by the Medical Ethical Committee of the Màxima Medical Center and the Central Committee for Human Research (CCMO). The

MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The results will be presented at conferences and published in peer-reviewed journals.

Trial registration number: NL9480 (www.trialregister.nl)

INTRODUCTION

In the Neonatal Intensive Care Unit (NICU), cardiorespiratory monitoring is crucial to assess clinical condition and to timely detect and treat frequently occurring cardio-respiratory events to prevent morbidity and mortality.(1, 2) To date, this is performed by measuring the electrocardiogram (ECG) and chest impedance (CI) with three wired adhesive electrodes. CI measures variation in electrical impedance across the chest during respiration caused by changes in lung aeration and chest wall movement. These techniques provide continuous monitoring of heart rate (HR), respiratory rate (RR), and breathing pattern. However, as CI measures respiration indirectly, adequate detection of breathing cycles and apnea may not always be optimal.(3)

With transcutaneous electromyography of the diaphragm (dEMG) breathing effort can be recorded directly by measuring the electrical activity of this main respiratory muscle. To date, this technique also uses three adhesive electrodes and provides information on respiration and HR. Studies have shown its feasibility in the NICU-setting.(4)

The use of adhesive electrodes is restricted in infants with a postmenstrual age <26 weeks in fear of skin damage.(5) Moreover, electrode removal may cause discomfort. Furthermore, the wires attached to the electrodes restrict movements of the infant and may hinder parent-infant interaction, nursing and kangaroo care. Restrictions in kangaroo care may impact patient outcome as it has been associated with beneficial effects such as decreased mortality, decreased

risk of severe infection/sepsis and hypothermia, and increased likelihood of exclusive breast feeding.(6, 7) All things considered, it is important to find alternatives for using wired adhesive electrodes.

In the past years, several wireless wearable sensors have been developed to measure various parameters in neonates such as ECG, HR, RR, peripheral oxygen saturation and (skin) temperature.(8-14) Recently, a novel wireless and non-adhesive sensor belt (Bambi® belt, Bambi B.V., Eindhoven, the Netherlands) was developed for neonatal use that measures ECG and respiration based on the dEMG technique. A recent pilot study showed that measuring HR and RR with this belt in preterm infants is feasible and that the measured HR and RR trend was similar to ECG and CI.(15) However, before replacing the current techniques using adhesive wired electrodes with the non-adhesive sensor belt, a larger study is required to demonstrate the non-inferiority of this belt as an alternative cardiorespiratory monitor. In this study, we compare the monitoring performance of the Bambi® belt to ECG and CI and hypothesize that the performance of the belt is non-inferior to the current monitoring techniques.

METHODS

Study design

This multi-center paired non-inferiority study will be performed in the NICU of Máxima Medical Center (MMC) in Veldhoven and the Emma Children's Hospital of the Amsterdam University Medical Centre (AmsterdamUMC), both located in the Netherlands. Each patient will be simultaneously measured with the belt and ECG/CI (paired design). To compare the devices, a non-inferiority/equivalence framework will be used. Here, equivalence is defined as the limit of agreement of the HR/RR between the belt and ECG/CI being within prespecified margins (see Table 1 for the margins). Non-inferiority is defined as the performance of clinical event detection and quality criteria not being worse than prespecified margins.

Study population

Preterm and term infants being routinely monitored with the standard cardiorespiratory monitor (Intellivue MP90, Philips Healthcare, Eindhoven, The Netherlands) are included in the study. To ensure a representative sample of the target population, infants in different age groups will be included. Infants with chest skin lesions, congenital anomalies, and other scenario's preventing belt placement, such as (effects of) surgery or wrap for therapeutic hypothermia, will be excluded.

Primary outcome

As HR-monitoring is clinically most relied upon and both ECG and dEMG provide the HR by measuring cardiac electrical activity, while CI and dEMG measure respiration with a different technique, the HR is considered the primary outcome.(3, 16) This will be studied with three criteria, which will be compared to the prespecified margins in Table 1. 1) Reliable monitoring performance through second-to-second HR measurement agreement in terms of differences in measured HR between the belt and the ECG/CI monitoring. 2) The detection of a composite cardiac event consisting of bradycardia (HR < 100 beats per minute for at least five seconds)(17) and tachycardia (HR > 180 beats per minute for at least ten seconds)(18) between the belt and the ECG measured with adhesive electrodes. The minimal duration of a bradycardia or tachycardia will prevent the inclusion of technical errors (short drops or increases in the HR) in our analysis and is lower for bradycardia compared to tachycardia as bradycardias are shorter events.(1) The thresholds are empirically chosen to detect all low and high HR-values. 3) Non-inferior quality (percentage of time with HR recordings without data loss).

Moreover, we will perform subgroup analyses to investigate whether the HR measurement performance is consistent under different clinical activities (e.g. kangaroo care, feeding) and in the different age groups.

Secondary outcomes

The secondary outcome is the measured RR. This will be studied using the following three criteria, which will be compared to the prespecified margins in Table 1:

1) Comparing the trend in RR values provided by the belt and CI, based on the difference in the 10-minute moving averages. The RR-trend is studied as this is used in the clinical practice to detect for example increases in RR over time as a marker of clinical deterioration of a patient.(3) Since CI is widely used for neonatal respiratory monitoring, it is used as the reference technique.
2) Next to comparing the RR-trend, the ability to detect apnea and tachypnea is studied as the detection of these critical respiratory events based on RR is another purpose of the respiratory monitoring. Clinically relevant apneas are considered when indicated by a RR < 20 breaths per minute measured with CI for at least 10 seconds, associated with a desaturation (arterial oxygen saturation as measured by pulse oximetry (SpO_2) <80% for at least 10 seconds) and/or bradycardia (HR <100 beats per minute for at least five seconds) (objective apnea measurement).(17) A RR<20 breaths per minute is chosen for the apnea definition as we solely use the numerical RR-values, because despite the two different measurement techniques this endpoint is equal, and to capture all periods of low breathing frequency.

   Tachypnea is defined as a prolonged period of the averaged (moving average with a window size of 10 minutes) RR >60 breaths per minute and >100 breaths per minute (approximately two times the average normal RR).(19) To cover short and long periods of tachypnea, 3 different durations are studied (30 seconds, 60 seconds, and 10 minutes).

3) Calculating the percentage of time with reliable respiratory monitoring (without data loss and with an acceptable signal-to-noise ratio).

Data collection

The following basic characteristics and demographic information will be collected at the baseline of the study: gestational age, birth weight, gender, age and weight at day of measurement, relevant medical status (respiratory support, medication and underlying illness during measurement), chest circumference, nipple distance, skin type at study start by visual inspection (normal, dry, flaky, oily, moist, other).

Sample size calculation

A power calculation is performed for the primary outcome using data collected in a previous study.(15) Among the three criteria, criteria 1 needs the largest sample size and is used for our study. This resulted in 39 required infants to achieve 80% power with an overall 5% type I error with a Bonferroni correction (details in the Statistical Analysis Plan (SAP) in the online supplement).

In addition, an interim analysis will be performed as the power calculation was based on the previous study and recruitment of infants without being able to answer research questions is unethical.(20) This will be performed after including 1/3th of the infants for sample size adaption using the method of Mehta and Pocock.(21) If the conditional power falls within the pre-defined "promising zone", the sample size will be increased to an upper limit of 52 infants. Otherwise, the study will proceed with the original sample size. To ensure that a representative sample of the age distribution of infants at a NICU, infants in different postmenstrual age groups will be recruited with the same proportions as in the target population (minimally 10 infants <30 weeks, between 30-32 weeks and >32 weeks).

Study procedures

The Bambi® belt system is a non CE-certified medical device, designed for wireless cardiorespiratory monitoring of (pre)term infants in a hospital environment. All included infants will be monitored with the belt in addition to standard ECG/CI for 24 hours to obtain representative clinical scenarios throughout the entire day. The measurement set-up is visualised in Figure 1 and consists of 1) dEMG measurement with the belt and 2) the extraction of patient monitor data.

In the belt, three dry electrodes are incorporated (Figure 2). When placing the belt at the height of the diaphragm, the outer two electrodes are in the nipple line and the middle electrode is in line with the sternum. The three ECG/CI electrodes are attached at the original location without hindering belt placement. The measured electrical signal of the diaphragm with the belt is wirelessly transmitted to the Receiver Module (REM) by the Sensor Module (SEM). The REM processes the dEMG signal to obtain the ECG and respiration signal (averaged diaphragmatic activity). An inbuilt algorithm provides the HR and RR out of the ECG and respiration signal respectively. This data is transported to a bedside computer. The data from the patient monitor (ECG, HR, RR, and SpO_2) is extracted from the bedside monitor using an isolated cable and is also transported to the bedside computer.

The belt data from the REM and patient monitor are recorded and synchronised using a dedicated software package (Polybench, Applied Biosignals, Weener, Germany) on a personal bedside computer. Data is recorded at a sample rate of 1 to 500 Hz for rate and waveform data respectively. The bedside software also provides the possibility to make measurements annotations by nurses and researchers during data recording, such as re-positioning of the infant, nursing and kangaroo care.

During the study, daily routine care proceeds as usual. The location of the belt is regularly checked and if necessary repositioned (similar to the clinical practice). Notifications are visualised when contact between skin and the belt is lost (Leads off) or when there is no connection between the SEM and REM (Bluetooth Loss Error). In case of the first notification, the belt may be repositioned, while in case of Bluetooth loss the battery level of the SEM or blocking of this sensor (e.g. by an arm) are checked.

Preferably, the belt stays in place during the study. However, the belt can be removed during diagnostic imaging, patient handling, or in case of skin irritation at the belt location. The reason for removal will be annotated. If the belt is removed, the medical staff, parents and one of the dedicated researchers will decide together if the belt can be re-applied.

Recruitment

Parents of all eligible infants are approached for consent to obtain a sample as heterogeneous and representative as possible. Preferably, infants are included as soon as possible after birth. During the 24 hours, the study can be terminated if requested by parents or the treating physicians. In case of withdrawal of a subject, an extra subject will be included.

Safety

Being a medical device study, this study was classified as a moderate risk.(22) A specified monitor plan for the study is made based on risk-classification.

STATISTICAL ANALYSIS

A detailed SAP can be found in the online supplement. Unless otherwise specified, all hypothesis tests are two-sided with a significance level of 0.05. All statistical analyses will be

performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

The non-inferiority/equivalence margins based on expert opinions (survey send to neonatologists of different NICU's in the Netherlands) and literature (4, 23, 24) are described in Table 1. In the different subparagraphs we refer to this table.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***Prespecified margins#***
LOA of second-to-second HR differences	± 8 bpm
LOA of RR trend differences	± 15 brpm
Sensitivity of brady-/tachycardia detection	90%˚
PPV of brady-/tachycardia detection	90%˚
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data with "Leads off" or "Bluetooth Loss Error" in the belt.

The prespecified margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

˚Note: all missed bradycardias are checked for clinical relevance by two independent experts.

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO_2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac interference, all values for PPV for apnea/tachypnea are acceptable. Interpretations will be made

Summary and descriptive statistics

Categorical data will be summarized by numbers of counts and percentages. Continuous data will be summarized by mean, standard deviation if data is normal and median, interquartile range (IQR) if data is skewed. Minimum and maximum values will also be presented for continuous data when appropriate.

Statistical analysis of the primary outcome

Criteria 1: agreement in HR

To investigate the equivalence of HR measurement between the belt and ECG, we will fit a linear mixed model to the second-to-second HR difference between both. With this model, the 95% limits of agreement (Bland-Altman analysis) will be derived. The two-one-sided tests (TOST) with a multiplicity corrected alpha of 0.0167 and the prespecified margin ($\pm$ 8bpm) will test equivalence between the two devices. In addition, based on a bivariate heteroscedastic model fitted to HR segments of a prespecified length, additional performance measures will be calculated as sensitivity analyses (details in SAP).

Criteria 2: cardiac event detection

For HR monitoring, we also consider the detection of bradycardia and tachycardia. We will estimate the sensitivity and the positive predictive value (PPV) of the belt using the patient monitor data as the ground truth and perform a non-inferiority test with an alpha of 0.0167. The non-inferiority margin for the sensitivity and PPV are listed in Table 1. In case of missed bradycardias, one independent expert per center will qualify the safety and clinical consequences of each missing event by answering the same questions per figure containing the discrepancy in HR and the ECG-signals measured with CI and the belt. These figures will be blinded and thus it will be unknown which signal corresponds CI or the belt.

Criteria 3: signal quality

The quality of the investigational device will be quantified based on the percentage of time during the 24-hour period it produces any reading (percentage without data loss due to "Leads off" or "Bluetooth Loss Error") and the percentage in time it produces a good-quality-reading (percentage of robust data) for the HR and RR, respectively. For the HR non-robust data can be caused by bad connection (suboptimal Bluetooth or skin-electrode connection). These criteria are built-in in the belt algorithm and therefore this data is automatically labeled. Hypothesis testing will be used to establish the non-inferiority of this "uptime" percentage (percentage without data loss and percentage of robust data) of the belt.

For the RR, the uptime percentage is also categorized as a) data readings without data loss and b) robust data readings, i.e. readings without unrealistic (e.g. negative) values. Signal quality is only analyzed for the belt. However, these results are compared to prespecified margins, described in Table 1. As the HR monitored with CI is accurate and nearly continuous, while the RR is less relied upon and may be unreliable, the prespecified margin for the RR is lower than for the HR.

Statistical analysis of secondary outcomes

Secondary analyses, based on the same statistical methods for the criteria of the primary outcome, include all secondary endpoints (apnea and tachypnea detection, RR trend analysis (see SAP)) and evaluation during different scenarios.

ETHICS AND DISSEMINATION

The Medical Ethical Committee of the MMC (W21.042) and the Central Committee for Human Research in the Netherlands (CCMO, CCMO21/0167/PP) approved the study protocol (Version 2, 19th of May 2021). Local feasibility at the AmsterdamUMC was approved by the Medical Ethical Committee of the AMC (2021_146**).** This study was registered in the Dutch Trial Register (https://www.trialregister.nl, NL9480). Regarding patient safety, no belt related events

were observed in the pilot study and are therefore unexpected. Moreover, as every patient is monitored with ECG/CI and the belt, safety is guaranteed in case of missing belt data. The SAP will be used for the analyses. The results will be published in peer-reviewed journals and presented at future congresses.

The MMC started patient recruitment in July and the AmsterdamUMC in August 2021. The duration of this study will be approximately seven months.

PATIENT AND PUBLIC INVOLVEMENT STATEMENT

Patients were included in this study after obtaining parental informed consent. The patients could not be involved in the design, recruitment, conduction and dissemination of results of this study. Neither could we ask the burden of the study. The outcome measures were developed by combining clinical and statistical knowledge to ensure a SAP that enables confirmation of non-inferiority of the belt compared to ECG/CI.

REFERENCES

1. Joshi R, van Straaten H, van de Mortel H, et al. Does the architectural layout of a NICU affect alarm pressure? A comparative clinical audit of a single-family room and an open bay area NICU using a retrospective study design. *BMJ Open* 2018;8(6):e022813. doi: 10.1136/bmjopen-2018-022813

2. Poets CF, Roberts RS, Schmidt B, Whyte RK, Asztalos E V, et al. Association Between Intermittent Hypoxemia or Bradycardia and Late Death or Disability in Extremely Preterm Infants. *JAMA*. (2015) 314(6):595–603. https://doi.org/10.1001/jama.2015.88413.

3. Di Fiore JM. Neonatal cardiorespiratory monitoring techniques. *Semin Neonatol* 2004;9(3):195-203. doi: 10.1016/j.siny.2003.11.009

4. Kraaijenga JV, Hutten GJ, de Jongh FH, et al. Transcutaneous electromyography of the diaphragm: A cardio-respiratory monitor for preterm infants. *Pediatr Pulmonol* 2015;50(9):889-95. doi: 10.1002/ppul.23116

5. Csoma ZR, Meszes A, Abraham R, et al. Iatrogenic Skin Disorders and Related Factors in Newborn Infants. *Pediatr Dermatol* 2016;33(5):543-8. doi: 10.1111/pde.12960

6. Boundy EO, Dastjerdi R, Spiegelman D, et al. Kangaroo Mother Care and Neonatal Outcomes: A Meta-analysis. *Pediatrics* 2016;137(1). doi: 10.1542/peds.2015-2238

7. Conde-Agudelo A, Diaz-Rossello JL. Kangaroo mother care to reduce morbidity and mortality in low birthweight infants. *Cochrane Database Syst Rev* 2016(8). doi: https://doi.org/10.1002/14651858.CD002771.pub4

8. Chung HU, Rwei AY, Hourlier-Fargette A, et al. Skin-interfaced biosensors for advanced wireless physiological monitoring in neonatal and pediatric intensive-care units. *Nat Med* 2020;26(3):418-29. doi: https://doi.org/10.1038/s41591-020-0792-9

9. Henry C, Shipley L, Ward C, et al. Accurate neonatal heart rate monitoring using a new wireless, cap mounted device. *Acta Paediatr* 2021;110(1):72-8. doi: https://doi.org/10.1111/apa.15303

10. Rimet Y, Brusquet Y, Ronayette D, et al. Surveillance of infants at risk of apparent life threatening events (ALTE) with the BBA bootee: a wearable multiparameter monitor. *Annu Int Conf IEEE Eng Med Biol Soc* 2007;2007:4997-5000. doi: 10.1109/IEMBS.2007.4353462

11. Chen H, Bao S, Lu C, et al. Design of an Integrated Wearable Multi-Sensor Platform Based on Flexible Materials for Neonatal Monitoring. *IEEE* 2020:23732-47. doi: 10.1109/ACCESS.2020.2970469

12. Chen W, Dols S, Bambang Oetomo S, et al. Monitoring body temperature of newborn infants at neonatal intensive care units using wearable sensors. *BodyNets '10* 2010. doi: https://doi.org/10.1145/2221924.2221960

13. Linti C, Horter H, Osterreicher P, et al. Sensory baby vest for the monitoring of infants. *IEEE* 2006. doi: 10.1109/BSN.2006.49

14. Pike H, Eilevstjonn J, Bjorland P, et al. Heart rate detection properties of dry-electrode ECG compared to conventional 3-lead gel-electrode ECG in newborns. *BMC Res Notes* 2021;14(1):166. doi: 10.1186/s13104-021-05576-x

15. Scholten AWJ, van Leuteren RW, de Waal CG, de Jongh FH, van Kaam AH, Hutten GJ. Feasibility of wireless cardiorespiratory monitoring with dry electrodes incorporated in a belt in preterm infants. *Physiological Measurement* 2022. doi: 10.1088/1361-6579/ac69a9

16. van Leuteren RW, Hutten GJ, de Waal CG, et al. Processing transcutaneous electromyography measurements of respiratory muscles, a review of analysis techniques. *Journal of Electromyography and Kinesiology* 2019:176-86. doi: 10.1016/j.jelekin.2019.07.014

17. Lee H, Rusin CG, Lake DE, et al. A new algorithm for detecting central apnea in neonates. *Physiol Meas* 2012;33(1):1-17. doi: 10.1088/0967-3334/33/1/1

18. Li T, Matsushima M, Timpson W, et al. Epidemiology of patient monitoring alarms in the neonatal intensive care unit. *J Perinatol* 2018;38(8):1030-8. doi: 10.1038/s41372-018-0095-x

19. Reuter S, Moser C, Baack M. Respiratory distress in the newborn. *Pediatr Rev* 2014;35(10):417-28. doi: 10.1542/pir.35-10-417

20. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

21. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

22. Central Committee on Research Involving Human Subjects.

23. Anton O, Fernandez R, Rendon-Morales E, et al. Heart Rate Monitoring in Newborn Babies: A Systematic Review. *Neonatology* 2019;116(3):199-210. doi: 10.1159/000499675

24. Breteler MJM, KleinJan EJ, Dohmen DAJ, et al. Vital Signs Monitoring with Wearable Sensors in High-risk Surgical Patients: A Clinical Validation Study. *Anesthesiology* 2020;132(3):424-39. doi: 10.1097/ALN.0000000000003029

AUTHOR CONTRIBUTIONS

AS, HN, MV, RL, FJ, AK, JH conceptualized the study. ZZ and EH made the statistical analysis plan, which was reviewed by all authors. AS wrote the first version of this manuscript. All authors contributed to the final draft of the manuscript.

FUNDING STATEMENT

This work was supported by the Louise Vehmeijer foundation.

COMPETING INTERESTS STATEMENT

Bambi Medical B.V. supports the study by a financial grant and use of equipment free of charge. Data collection, analysis, interpretation, and reporting will be done independent of Bambi Medical B.V..

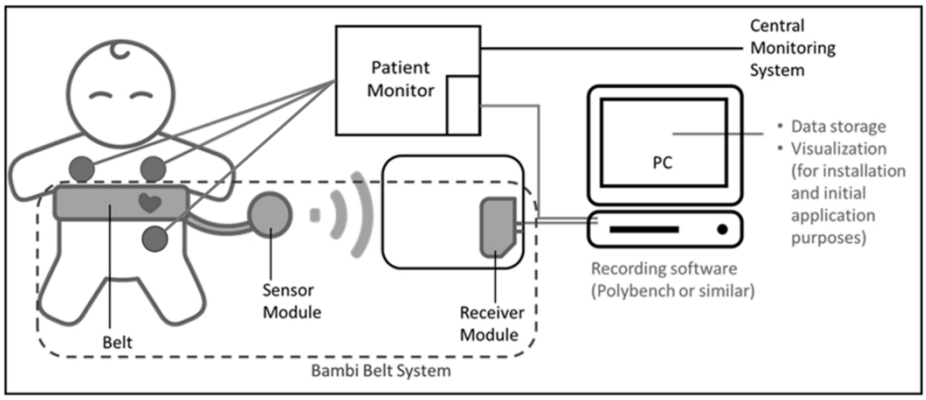
HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- When non-inferiority of the Bambi® belt compared to the current cardiorespiratory monitor is confirmed, the belt could be used as a wireless and skin-friendly alternative.

FIGURES

FIGURE 1 - The measurement set-up. The adhesive electrodes used for standard cardiorespiratory monitoring are attached at the original location, visualised by the three grey dots. The diaphragm activity measured with the Bambi® belt is wirelessly transmitted with the Sensor Module to the Receiver Module where the data is processed to obtain an electrocardiogram and respiration waveform (and heart rate and respiratory rate). This data and the data measured with the patient monitor are transported to a personal bedside computer with Polybench software to synchronise and record these signals.

FIGURE 2 – The Bambi® belt is a wireless non-adhesive belt designed for cardiorespiratory monitoring of (pre)term infants. The three dry electrodes (2) measure electrical activity of the diaphragm via transcutaneous electromyography. This data is wirelessly transmitted with the sensor module (1) to a receiver module that processes the diaphragm activity to obtain the ECG, respiration signal, heart rate and respiratory rate.



170x80mm (300 x 300 DPI)

170x39mm (300 x 300 DPI)

ONLINE SUPPLEMENT – STATISTICAL ANALYSIS PLAN

PURPOSE OF THE STATISTICAL ANALYSIS PLAN

The purpose of this statistical analysis plan (SAP) is to document technical and detailed specification for the analysis of data collected in the Bambi belt monitoring performance study. The SAP has been written based on information contained in study protocol, dated 12th April 2021 before any data collection had taken place. It is prepared in compliance with the International Council on Harmonization (ICH) E9.

This SAP will be the guiding document for the analyses that will be conducted. Results of the analyses described in this SAP will be included in the Clinical Study Report (CSR). Any post hoc or unplanned analyses performed to provide results for inclusion in the CSR, but not identified in the prospective SAP will be identified in the given report. Additionally, the planned analyses of the primary aims will be included in future manuscripts. All the aims and research questions will be presented as an addendum as well.

OVERVIEW AND DESCRIPTION OF THE STUDY

Study design

The study is a multi-center, paired design, clinical monitoring device measurement comparison study. The *investigational device* under consideration is the Bambi® belt monitoring system (using dry electrodes). The current standard device of cardiorespiratory monitoring through adhesive electrodes is considered as the clinical reference standard and thereafter referred to as the *reference device/method*. The Bambi® belt monitoring system will be used on infants by trained nurses in the neonatal intensive care units (NICU's) for continuous 24 hours monitoring in addition to the routine monitoring with the reference device on the same patients. Infants admitted to NICU's of the the Emma Children's Hospital of the Amsterdam University Medical

Centre (Amsterdam UMC) or Maxima Medical Center (MMC) will be measured at the earliest suitable moment for clinical practice without interfering with infants' routine cycles.

Randomization and blinding

No randomization is required for the paired design since both monitoring devices will be used on the same patient at the same time. Blinding is also not possible since both the measurement protocol and algorithmic characteristic differ substantially.

Framework

The goal of this study is to establish the agreement between the investigational device and the reference device. Unlike the traditional difference-based tests, non-inferiority and equivalence techniques provide a better alternative for demonstrating the similarity between the two measurement methods. Thus, we have adopted the non-inferiority/equivalence trial framework for this primary objective of this study. This study considers three hypotheses (H_0 denotes the null hypothesis and H_A denotes the alternative hypothesis) for the first two primary outcomes:

1. Primary outcome, criterion 1: Heart rate measurement (second-by-second measurement)

H_0: The absolute difference between the investigational device and the reference device is larger than the prespecified equivalence margin.

H_A: The absolute difference between the investigational device and the reference device is within the prespecified equivalence margin.

2. Primary outcome, criterion 2: Brady-/tachy-cardia event detection

H_0: The composite cardiac event detection performances in terms of sensitivity and positive predictive value (PPV) based on the investigational device with respect to the reference device is less than the prespecified non-inferiority margin.

H_A: The composite cardiac event detection performances in terms of sensitivity and PPV based on the investigational device with respect to the reference device is greater or equal to the prespecified non-inferiority margin.

3. Primary outcome, criterion 3: Reliable reading (percentage of the time)

H_0: The percentage of the time the investigational device produces reliable readings is less than the prespecified non-inferiority margin.

H_A: The percentage of the time the investigational device produces reliable readings is greater or equal to the prespecified non-inferiority margin.

Statistical interim analysis and stopping guidance

One interim analysis for sample size adaptation will be performed. That is, we will start with a certain sample size commitment which will be increased at the interim analysis in case the results obtained are reasonably promising. The interim analysis will be conducted after the prospectively recruited participant's number reaches one-third of the planned sample size. Conditional power will be calculated for the analyses of the primary endpoints and compared to the boundary values of the conditional power for the promising zones (1, 2). In case the conditional power calculated at the interim analysis does fall inside the promising zone, the sample size will be increased to a predetermined limit. On the other hand, if the calculated conditional power is outside the promising zone, the study will proceed with the original sample size. Therefore, no early stopping rule is entailed in this study. Furthermore, a conventional

final analysis will be used without altering the level of type I error, since the promising zone is defined as a set that ensures the type I error to be preserved conservatively for the final analysis.

Study data

The following infant characteristics will be collected at baseline:

- Gestational age
- Postmenstrual age
- Gender
- Birth weight
- Weight at enrollment
- Ethnicity (derived from the electronic patient record or by asking the parents)
- Chest circumference
- Nipple distance
- Skin condition and abnormality

During the monitor study period, the following information will be measured:

- Clinical event
- SpO2: Arterial oxygen saturation as measured by pulse oximetry
- Medical status:
    - Ventilation support
    - Reports of medication and illness during the measurement
- Lead status: Indicates whether at least one lead was off
- Bluetooth link quality
- Activities
    - Kangaroo care
    - Nurse care

- Feeding
- Medical Procedure

- Belt status
  - Moved: the belt is being moved
  - Open: the belt is removed from the patient
- Patient position
  - Unknown
  - Lying prone
  - Lying supine
  - Lying on the left side
  - Lying on the right side

STATISTICAL ANALYSIS

Based on the collected information described above, the following total of variables will be derived:

- 10, 30, and 60 minutes moving average of the heart rate, and respiratory rate measured by both the investigational device and the reference device.
- Premature birth:
  - Premature (gestational age < 37 weeks)
  - Normal (gestational age >= 37 weeks)
- Desaturation: SpO2 < 80% for at least 10 consecutive seconds
- Heart rate status (investigational and reference device):
  - Normal
  - Tachycardia (heart rate > 180 for at least 10 consecutive seconds)
  - Bradycardia (heart rate < 100 for at least 5 consecutive seconds)

- Respiration status (investigational and reference device):
    - Apnea (according to standard clinical definitions)
    - Tachypnea (respiratory rate >60 and >100 for 30 seconds, 1 minute, and 10 consecutive minutes in stationary signal)
- Measurement quality:
    - No anomalies
    - Poor data link: Bluetooth link is poor but data is still received
    - Unreliable data: One or more lead off, or no Bluetooth connection (Bluetooth Loss Error, BLE)

Summary and descriptive statistics

Categorical data will be summarized by numbers and percentages. Continuous data will be summarized by mean, standard deviation if data are normal and median, interquartile range (IQR) if data are skewed. Minimum and maximum values will also be presented for continuous data. Tests of statistical significance will not be undertaken for baseline characteristics; rather the clinical importance of any imbalance will be noted.

A CONSORT flow diagram (example in Figure 1) will be used to summarize the number of infants who were:

- Assessed for eligibility at the screening
    - Eligible at screening
    - Ineligible at screening (with reasons)
- Eligible and enrolled
- Eligible but not enrolled
- Enrolled but did not receive any / sufficient measurements

- Discontinued

- Included in the analysis
- Excluded from the analysis

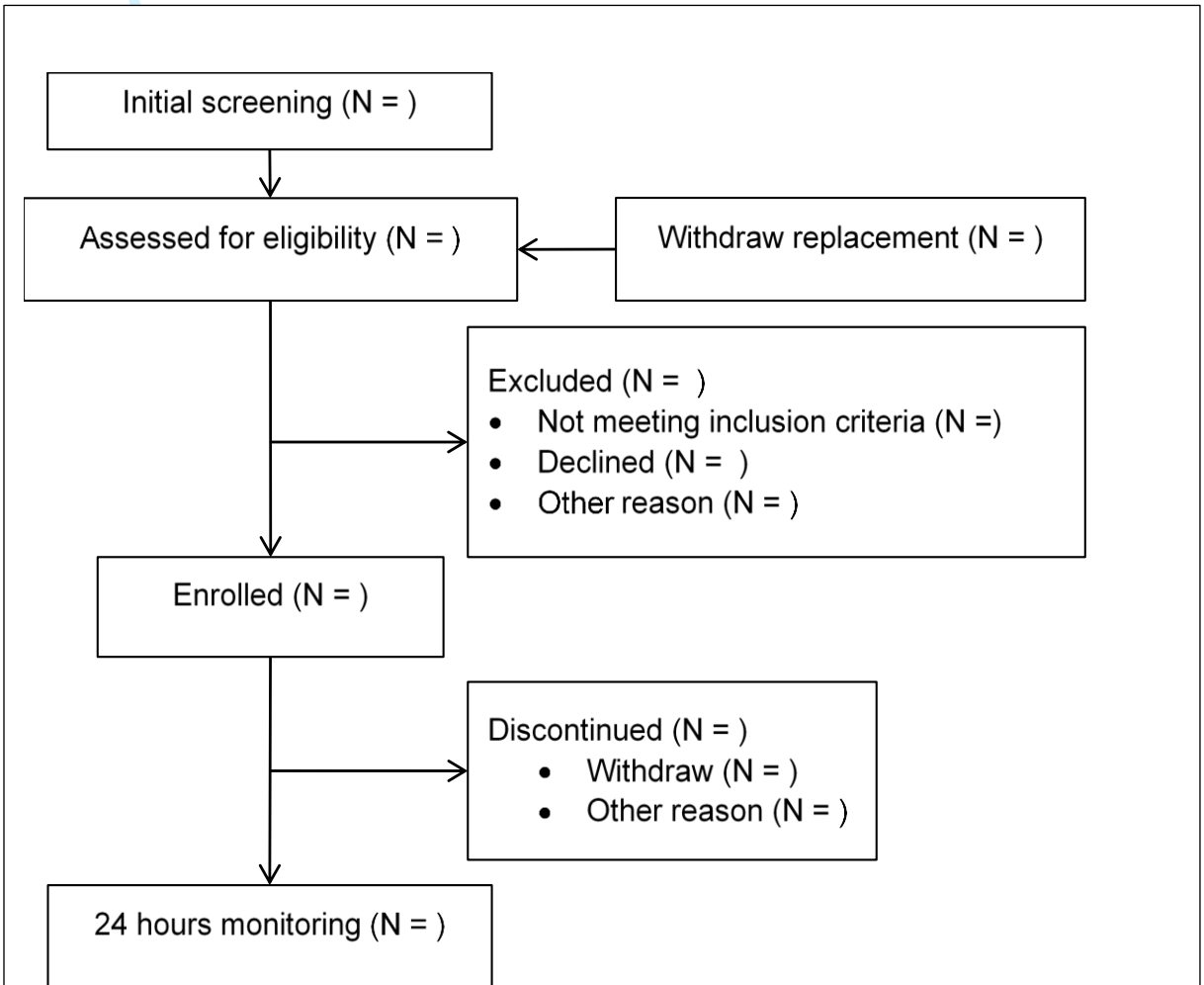


FIGURE 1 – Example of a CONSORT flow diagram

Analysis methods

Primary outcome, criterion 1: Heart rate measurement

To investigate and verify the equivalence of heart rate measurement between the investigational device and the reference device, we will fit a linear mixed model to the second-to-second heart rate difference between the two devices. Based on the estimates of the model, we will derive the 95% limits of agreement (3) as our main performance measure, known as the Bland-Alman analysis. The endpoints of the Bland-Altman 95% limits of agreement are the 2.5th percentile and 97.5th percentile for the distribution of the difference between paired measurements. We will calculate the $(1-\alpha/2)100\%$ confidence intervals of the percentiles according to Shieh (4), and conduct the two-one-sided t-tests (TOST) procedure with the prespecified equivalence margins (Table 1).

In addition, we will calculate the following performance measures to supplement the main analysis as sensitivity analyses to assess the agreement between the two devices from different aspects:

- The concordance correlation coefficient (5) and its variants
- Probability of Agreement (6) and Total Deviation Index (7)
- Coefficient of individual agreement (8)

These performance measures will be based on a bivariate heteroscedastic linear mixed-effects model fitted to each segment of the readings of a prespecified length from both devices. We will assume that measurements made with the two devices at the same time are correlated. Therefore, investigating the correlation between the two devices leads to the quantification of the degrees of agreement between them. Furthermore, we will consider the temporal correlations between measurements obtained with the same devices and the variabilities between different infants. Besides, we will start with a heteroscedastic model which does not assume equal variances for the two devices (namely, the measurement errors are not assumed

to be equal) and investigate the homogeneity of the measurement variabilities between the two devices. Baseline characteristics of the infants and records of activities (listed in the study data section) will be used as covariates in the model to partly explain the variabilities between the infants. We will use the stepwise model selection procedure based on the Bayesian Information Criteria (BIC) goodness-of-fit criteria.

Primary outcome, criterion 2: Brady-/tachycardia event detection

For brady-/tachycardia, the clinical event periods will be identified based on prespecified margins. We will investigate the non-inferiority of sensitivity and positive predictive values (PPV) of the event detected by the investigation device assuming that the reference device is the predicate device and compare both values to the prespecified non-inferiority margins (Table 1). For the calculation of the sensitivity, when the event period identified based on the investigational device overlaps with the event period identified by the reference device, it will be counted as a true positive case. This is to prevent the repeated signaling of events from the investigational device during a positive period identified by the reference device to inflate the number of true positives. The same applies to the reference device when it comes to the calculation of the PPV. That is, during an event period identified by the investigational device, multiple event periods identified by the reference device will only be counted as one true-positive case. Note that the true negative is ill-defined and will not be reported. Since true negatives are used in the calculation of specificity, specificity will not be reported either.

Primary outcome, criterion 3: Safety and Quality

Safety: The investigation of safety and tolerability is a multidimensional problem. Although we don't anticipate any specific adverse effects for the investigational device, new and unforeseeable effects are always possible. This background underlies the statistical difficulties

associated with the analytical evaluation of the safety and tolerability of the device. We will address the safety and tolerability implications by applying descriptive statistical methods to the data, supplemented by calculation of confidence intervals whenever this aids interpretation and make use of graphical presentations in which patterns of adverse events are displayed.

Quality: The quality of the investigational device will be quantified in terms of the point estimate and 95% confidence intervals based on the estimated percentages in time during the 24-hour period it produces reliable readings for heart rate and respiratory rate, respectively. Reliable readings are defined in the study protocol. The uptime percentages are the percentage of data loss and the percentage of robust data readings. For each outcome, hypothesis testing will be used to establish the non-inferiority of the uptime percentages of the investigational device considering a non-inferiority margin specified in Table 1. The uptime percentages will be estimated based on a Generalized Estimation Equations (GEE) model.

Missing data

To get an idea about the complexity of the missing data problem in the data and information about the location of the missing values, the missing data pattern will be evaluated and reported. We expect missing data in the primary outcomes measured by the investigational device to be the results of external causes such as the movement of the belt, signal losses, poor Bluetooth link qualities and so on. Therefore, it will be reasonable to assume that data are missing completely at random (MCAR). Formally, we will investigate the validity of such an assumption using Little's MCAR test. Furthermore, the availability of the data from the reference device (since it depends on a separate measurement system) provides us the opportunity to investigate whether the missingness is related to the underlying measurand. That is, whether the missing data mechanism is missing not at random (MNAR). This is rarely

possible in other types of studies. Nevertheless, considering the pair of bivariate measurements from the investigational and the reference device, we will investigate the assumption using the covariate-dependent missing (CDM) test proposed in Li (9). Note that CDM is usually considered as missing at random (MAR), we here simply exploit the advantage of the data from the reference device to test the dependencies between the missingness and the underlying measurand. Furthermore, we will use the CDM test on other covariates (excluding the reference device data) as well to test if the missingness is MAR.

In the case of MAR (i.e., CDM without measurements from reference device), list wise deletion can still be unbiased and will be used if the percentage of missingness is less than 5%. Otherwise, multiple imputations (MI) will be considered. We will not use the measurements from the reference device for the MI to avoid biasing the results towards the equivalence of the two devices. On the other hand, if the missingness is related to the measurand after taking into account all covariates, this indicates a potential problem of the measurement device, and a separate analysis will be carried out to investigate the associations between the missingness and the measurand.

For multiple imputations, we will use the fully conditional specification method. Unrealistic values (e.g., negative values for strictly positive variable) will be checked and corrected (e.g., using truncations). The imputation will be repeated at least 5 times and Rubin's rule will be used to combine estimates and standard errors from the imputed data.

Secondary analyses

If the sample size permits, we will perform subset analyses to explore the performances of the investigational device under different scenarios.

Subset analyses: primary endpoints

For each of the primary endpoints, we will consider additional exploratory analyses on the following subsets:

- During periods of a clinical event (e.g., apnea, bradycardia)
- During activities (e.g., Kangaroo care, feeding)
- During periods where the reference device's readings are stable
- Gestational age (e.g., preterm birth)
- Respiratory support (e.g., mechanical ventilation)

For these subsets, we will use the same model as the primary outcome to investigate the performances of the investigational device under various scenarios/activities of the infants. In case the subset does not contain enough data to fit the same model as the primary one, we will resort to a simpler model for case-by-case analyses.

Respiratory rate analysis

It is known the reference device does not provide point-by-point accurate measurement resulting in large variabilities (measurement errors) in the measured respiratory rates. The intended clinical use of the readings in the NICU thus consists of two different aspects:

1. The trend of the respiratory rates over time;
2. Signaling of potentially respiratory related clinical events (i.e. apnea related desaturation and/or bradycardia, and potentially disease related tachypnea);

For the first usage, we will apply the same analysis method as the one used for heart rate on the moving average of the respiratory rate. We will primarily focus on the 10 minutes moving average for the respiratory rate. Analysis of the 1 minute and 5 minutes moving averages will

be used as a sensitivity analysis to establish the robustness of the conclusions made for the 10 minutes moving average.

For apnea and tachypnea, respectively, the clinical event periods will be identified based on clinical definitions and the same methods as the brady-/tachycardia event detection will be used to compare the sensitivity and PPV to the prespecified non-inferiority limits (Table 1). However, it should be noted that since the reference device is known to have an unsatisfactory performance of apnea/tachypnea detection, cautions are needed to interpret the sensitivity and PPV as if the reference device is the truth.

Statistical software

All statistical analyses will be performed using R version 4.0 (the R Foundation for Statistical Computing; Vienna, Austria) and SAS software version 9.4 (SAS Institute Inc., Cary, NC, USA).

Non-inferiority/equivalence criteria

In Table 1 the non-inferiority/equivalence criteria for the primary and secondary outcomes are visualized.

Table 1 - The non-inferiority/equivalence margins for the primary and secondary outcomes

Endpoints	***Prespecified margins#***
LOA of second-to-second HR differences	±8 bpm
LOA of RR trend differences	±15 brpm
Sensitivity of brady-/tachycardia detection	90%˚
PPV of brady-/tachycardia detection	90%˚
Sensitivity of apnea/tachypnea detection	70%
PPV of apnea/tachypnea alarms	0-100%*
Data loss percentage	5%
Robust data percentage (HR)	90%
Robust data percentage (RR)	70%

LOA: limits of agreement, HR: heart rate, RR: respiratory rate, PPV: positive predictive value. Data loss is defined as the percentage of data with "Leads off" or "Bluetooth Loss Error" in the belt.

The prespecified margins are compared to confidence intervals with corresponding confidence levels (see SAP for more details).

˚Note: all missed bradycardias are checked for clinical relevance by two independent experts.

*Since the reference devices for apnea detection in the clinical practice are the peripheral oxygen saturation (SpO2) and electrocardiogram instead of the respiration signal and the performance for Chest Impedance to detect tachypnea is unsatisfactory due to the presence of cardiac

Sample size

Based on preliminary analysis of data collected in a feasibility study on a total of 13 infants with measurements from both the investigational device and the reference device, we were able to obtain preliminary information with regards to the characteristics of the primary endpoints

upon which we have formulated our sample size calculation.(submitted for publication, NICU AmsterdamUMC, 2021)

A detailed specification of the sample size calculation can be found in the sections below. In summary, for the monitor performance, 39 infants are needed to achieve 80% power with a 5% overall type I error with a Bonferroni correction for multiplicity. It is worth noting that no dropout was assumed during the sample size calculation. This is because we plan to include an extra infant in case of withdrawal of an infant to fulfil the required sample size. Infants who withdraw from the study will be followed up by one of the investigators and responsible medical staff to obtain detailed reasons behind the withdraw. Dropout rate for the monitor performance study is expected to be low, between 0-5%.

While the preplanned sample size is 39 infants, we will include an adaptive sample size re-estimation procedure as per the "promising zone" methodology of Mehta and Pocock (2) using the data from the first 1/3 infants. This procedure involves the evaluation of conditional power in the interim analysis, and if it were to fall in the prespecified "promising zone", the sample size will be increased, subject to a predetermined upper limit (52 infants) to increase the conditional power to 80%. The boundary of the conditional power for the "promising zone" is 0.36 and 0.8.

Monitoring study: Primary endpoints: Heart rate

For the sample size calculation, we will assume the measured heart rate difference D_{ij} between the investigational device and the reference device at time point j $(j = 1, \dots, m)$ on infant i $(i = 1, \dots, n)$ can be modelled as:

$$D_{ij} = d + a_i + e_{ij}$$

where d is the overall difference, a_i is a random effect with $a_i \sim N(0, \sigma_a^2)$, and $e_{ij} \sim N(0, \sigma_e^2)$ is the random error independent of a_i. Though, we considered a bivariate mixed-effects model for our analysis, the variance component model for the difference can be derived from the bivariate mixed-effects model, therefore we will use this variance component model for the sample size calculation. The variance of the difference will be estimated from the aforementioned model via $\hat{\sigma}_d^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2$. Here $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ is the estimator of the between-subject variability σ_a^2 and residual variability σ_e^2, respectively. The 95% limit of agreement (LOA) can be estimated as LOA = $\hat{d} \pm 1.96\,\hat{\sigma}_d$ with $\hat{d}$ and $\hat{\sigma}_d$ denotes the estimator of d and σ_d, respectively. The variance of the LOA estimator is $\text{var}\left(\hat{d} \pm 1.96\,\hat{\sigma}_d\right) = \text{var}\left(\hat{d}\right) + 1.96^2\text{var}(\hat{\sigma}_d)$ ($\hat{d}$ and $\hat{\sigma}_d$ is asymptotically independent). Since for $\hat{\sigma}_d^2 = \hat{\sigma}_a^2 + \hat{\sigma}_e^2$, we have $\text{var}(\hat{\sigma}_d^2) = \text{var}(\hat{\sigma}_a^2) + \text{var}(\hat{\sigma}_e^2) + \text{cov}(\hat{\sigma}_a^2, \hat{\sigma}_e^2)$. Furthermore, each term on the right-hand side (assuming m is large) is given by:

$$var\,(\hat{\sigma}_a^2) = \frac{2}{m^2}\left[\frac{(m\sigma_a^2 + \sigma_e^2)^2}{n-1} + \frac{\sigma_e^4}{n(m-1)}\right] \approx \frac{2\sigma_a^4}{n-1}, var\,(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{n(m-1)+2} \approx 0,$$

$$cov(\hat{\sigma}_a^2, \hat{\sigma}_e^2) = -\frac{2\sigma_e^4}{nm(m-1)} \approx 0;$$

This leads to $\text{var}\left(\hat{\sigma}_d{}^2\right) \approx 2\sigma_a{}^2/(n-1)$. Therefore, by the delta method, we have $\text{var}(\hat{\sigma}_d) = \frac{1}{4\sigma_d{}^2} var(\hat{\sigma}_d{}^2) = \frac{\sigma_a{}^4}{2(n-1)\sigma_d{}^2}$. According to Lu et al. (10), the power for the TOST is given by:

$$1 - \beta = 1 - T_{n-1}\left(t_{1-\frac{\alpha}{2}}, \frac{\delta - d - 1.96\sigma_d}{se_{LOA}}\right) - T_{n-1}\left(t_{1-\frac{\alpha}{2}}, \frac{\delta + d - 1.96\sigma_d}{se_{LOA}}\right)$$

where α, β denotes type I and type II error respectively, δ is the predefined limit, $se_{LOA} \approx \sqrt{\frac{{\sigma_d}^2}{n} + \frac{1.96^2{\sigma_a}^4}{2(n-1){\sigma_d}^2}}$ is the standard error of the LOA estimate calculated according to the variance component model, and $T_{n-1}(\cdot, \tau)$ denotes the cumulative distribution function of a non-central Student's t-distribution with n-1 degrees of freedom, and non-centrality parameter τ.

For a 5% overall type I error rate, with a multiplicity correction factor of 3, and 80% power, the minimum sample size required is calculated at $n = 39$, for $d = -0.5$, $\sigma_a = 0.3$, and $\sigma_d = 3$.

Primary endpoints: Brady-/tachycardia event detection

Suppose the total number of true events is M and are 100% detected by the reference device. Assuming the true sensitivity is 95% for the investigational device, then a non-inferiority test using Z-test with normal approximation to the binomial distribution leads to a required M of 271 for a power of 80% and $\alpha = 0.05/3 \approx 0.01667$ assuming the detection between each event (conditioning on the event itself) is independent. Considering the incidence of bradycardia to be 1 event per hour per infant according to the preliminary analysis of data from the feasibility study, at least 12 infants are needed to satisfy the required M (assuming each infant is measured for 24 hours long). The calculation is the same for PPV if we assume the investigational device is the truth. Assuming an incidence rate of tachycardia of 1.5 per hour per infant according to the preliminary analysis of data from the feasibility study, the required sample size is 8. Note that in the aforementioned calculation, we assume that the event-detection performance of the investigational device is homogeneous (or independent) among infants. A

sensitivity/robustness investigation regarding the sample size for infant-specific heterogeneous performances was performed, with results from which we can see that with $n = 39$, we have more than 90% power to detect a heterogeneous performance scenario where 15% of the population would have sensitivities between 80% - 90% and less than 5% of the population have sensitivities less than 80%.

Primary endpoints: Safety and quality

Based on preliminary analysis of data collected in the feasibility study, we will assume that the overall probability of producing an erroneous reading at any time p_e is 2% and is constant across all participants. We will consider a non-inferiority test using normal approximation and a Z-test with the null hypothesis of $H_0: p_e > 0.05$ and the alternative hypothesis of $H_A: p_e \leq 0.05$. The required number of observations for a given type I error of 1.667% ($\approx 5\%/3$) to achieve 80% power is 376. Here the sample size 376 refers to 376 independent observations. Considering the large numbers of repeated measurements (more than 376) within each participant, we will have sufficient power for this non-inferiority test even with 1 participant. However, the assumption of independence can be too strong in the setting of our study. Therefore, if we would assume an AR(1)-type dependency with correlation parameter $\rho = 0.8$ between two measurements within a participant, the variance inflation factor (VIF) for the asymptotic variance of the GEE estimator $\hat{p}_e$ according to Pan is approximately (with the number of repeats $m = 376$):

$$1 + \frac{2\rho}{1-\rho} = 9$$

in the case of identity working correlation matrix when the true correlation has an AR(1) structure. To achieve the same power as the independent case calculated before, we need

$$var(\hat{p}_e) \coloneqq VIF\,\frac{p_e(1-p_e)}{nm} = \frac{p_e(1-p_e)}{m}$$

Thus, we can conclude that at least $n = \text{VIF} = 9$ participants will be needed to provide enough power for the non-inferiority test based on the GEE estimator using the identity working correlation matrix using the inverse proportionality between the required sample size and the variance of the estimator used in the Z-test. The same calculation can be carried out for the robust data percentages. It can be seen that only the number of repeats m will differ when the probabilities and the non-inferiority margins change while the VIF remains the same for the same value of the correlation parameter ρ. Among all settings, the largest m needed will be 718 when we assume the probability of producing robust data for respiratory rate is 75% with the corresponding non-inferiority margin equals to 70%. This number of repeats is still fully covered by the high-frequency measurements found in the study.

Protocol deviations and analysis sets

Definition of protocol deviations

Protocol deviations (PD) occurring during the study will be determined for all enrolled infants, mainly from the clinical database by either clinical and/or medical review processes.

The mapping of the protocol deviations from the clinical database to analysis will be performed as per Table 2:

Table 2 – The influence of protocol deviations on the statistical analysis plan (SAP)

Database label	*SAP*
Minor	Not required
Major	Important
Critical	Important
Clinical (a subset of Critical)	Important

Important protocol deviations are protocol deviations that might significantly affect the completeness, accuracy, and/or reliability of the study data or that might significantly affect a subject's rights, safety, or well-being.

Important protocol deviations may also be recorded as "Major" protocol deviations in the database, but will be presented only as important in the analysis output.

Important protocol deviations include:

- Infants that are included in the study despite not satisfying the eligibility criteria;
- Infants that develop exclusion criteria while on the study but not withdrawn;
- Infants being measured with operational human errors;
- Deviation from Good Clinical Practice (ICE E6)

Clinically Important protocol deviations are the protocol deviations marked as important in Table 2, which lead to the exclusion of a subject from the analysis set.

The following deviations will be identified and confirmed before the partial database lock for the final analysis.

- Important protocol deviations including
    - Deviations from the inclusion and exclusion criteria
    - Deviations post inclusion

Protocol deviations may be identified by the data managers, clinical and medical staff either by programmed validation checks or data listings/reports or manual verification of data sources. Some important/major protocol deviation criteria may be identified in the clinical database via biostatistical programs. Every important protocol deviation will be documented in the database whether identified through sites monitoring, medical review or programming.

REFERENCES

1. European Medicines Agency. ICH Topic E 9 – Statistical Principles for Clinical Trials. 2006:21-3.

2. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med* 2011;30(28):3267-84. doi: https://doi.org/10.1002/sim.4102

3. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8(2):135-60. doi: 10.1177/096228029900800204

4. Shieh G. The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. *BMC Med Res Methodol* 2018;18(1):45. doi: https://doi.org/10.1186/s12874-018-0505-y

5. Lin L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989:255-68.

6. Stevens NT, Steiner SH, MacKay RJ. Assessing agreement between two measurement systems: An alternative to the limits of agreement approach. *Stat Methods Med Res* 2017;26(6):2487-504. doi: 10.1177/0962280215601133

7. Lin L. I.-K. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* 2000:255-70.

8. Haber M, Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurements. *Stat Methods Med Res* 2008;17(2):151-69.

9. Li C. Little's Test of Missing Completely at Random. *The Stata Journal* 2013. p. 795-809. doi: 10.1177/1536867X1301300407

10. Lu MJ, Zhong WH, Liu YX, et al. Sample Size for Assessing Agreement between Two Methods of Measurement by Bland-Altman Method. *Int J Biostat* 2016;12(2). doi: 10.1515/ijb-2015-0039