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 primary outcome:

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-/tachycardia 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
    - Nursing 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:

- Second-to-second heart rate and 10, 30, and 60 minutes moving average of the 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

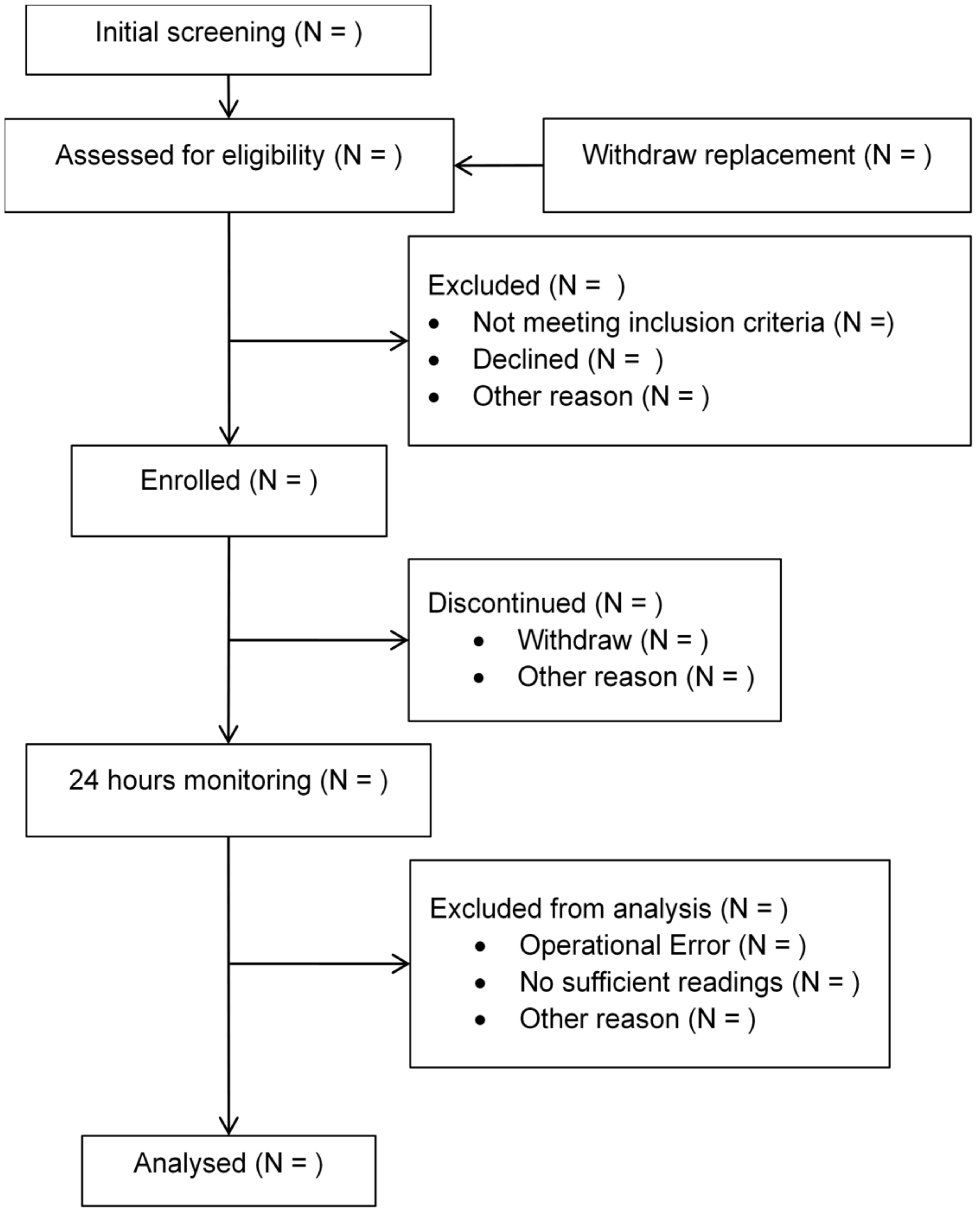

Figure 1 Example of a CONSORT flow diagram

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.

Scholten AWJ, *et al. BMJ Paediatrics Open* 2022; 6:e001430. doi: 10.1136/bmjpo-2022-001430

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 that 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, SAP: statistical analysis plan.

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 based on the results.

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.(10)

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.

Primary endpoint: 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 $\text{var}(\hat{\sigma}_d{}^2) \approx 2\sigma_a{}^2/(n-1)$. 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. (11), 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 endpoint: 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 endpoint: 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. 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

11. 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