Received: 17 January 2024 | Revised: 30 April 2024 | Accepted: 24 July 2024

DOI: 10.1111/birt.12858

ORIGINAL ARTICLE

BIRTH ISSUES IN PERINATAL CARE WILEY

# How well can the fetal heart rate baseline be assessed by intrapartum intermittent auscultation? An interrater reliability and agreement study

**Christina Hernandez Engelhart MMid**[1,2] | **Sophie Vanbelle PhD**[3] | **Pål Øian MD, PhD**[4] | **Aase Serine Devold Pay PhD**[2,5] | **Anne Kaasen PhD**[2] | **Ellen Blix PhD**[2]

[1]Norwegian Research Centre for Women's Health, Oslo University Hospital, Oslo, Norway

[2]Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway

[3]Department of Methodology and Statistics, Maastricht University, Maastricht, The Netherlands

[4]Department of Gynaecology and Obstetrics, University Hospital of North Norway, Tromsø, Norway

[5]Department of Gynaecology and Obstetrics, Bærum Hospital, Vestre Viken Hospital Trust, Bærum, Norway

**Correspondence**
Christina Hernandez Engelhart, Oslo Universitetssykehus, Rikshospitalet, Nasjonalt Senter for Kvinnehelseforskning, Postboks 4950 Nydalen, 0424 Oslo, Norway.
Email: cheng@oslomet.no

**Funding information**
The Norwegian Research Centre for Women's Health, Oslo University Hospital

## Abstract

**Background:** We aimed to examine the inter-reliability and agreement among midwives when assessing the fetal heart rate (FHR) using the handheld Doppler. The primary aim was to measure the reliability and agreement of FHR baseline (baseline) as beats per minute (bpm). The secondary aims were to measure fluctuations from the baseline, defined as increases and decreases, and classifications (normal or abnormal) of FHR soundtracks. This is the first interrater reliability and agreement study on intermittent auscultation (IA) to our knowledge.

**Methods:** The participant population consisted of 154 women in labor, from a mixed-risk population and admitted to hospital for intrapartum care. The rater population were 16 midwives from various maternity care settings in Norway. A total of 154 soundtracks were recorded with a handheld Doppler device, and the 16 raters assessed 1-min soundtracks once, through an online survey (Nettskjema). They assessed the baseline, FHR increase or decrease, and the FHR classification. The primary outcome, baseline, was measured with intraclass correlation coefficient (ICC). The secondary outcomes were measured with kappa and proportion of agreement.

**Results:** The interrater reliability for the baseline (bpm) was ICC(A,1) 0.74 (95% CI 0.69–0.78). On average, an absolute difference of 7.9 bpm (95% CI 7.3–8.5 bpm) was observed between pairs of raters.

**Conclusion:** Our results demonstrate an acceptable level of reliability and agreement in assessing the baseline using a handheld Doppler.

**KEYWORDS**

agreement, fetal monitoring, handheld Doppler device, intermittent auscultation, reliability

## 1 | INTRODUCTION

Intrapartum fetal heart rate (FHR) monitoring is a widely used clinical assessment of fetal well-being during labor and birth. The aim of the monitoring is to detect fetuses with signs of compromised oxygen supply and guide clinical decisions for intrapartum interventions.[1] FHR monitoring is based on the rationale that changes in the FHR can indicate compensation for excessive intrapartum stress due to low oxygen supply.[2] All fetuses are prone to changes in the oxygen supply during birth, mostly because of repetitive compression from uterine contractions. Most fetuses tolerate contractions well, if they are within the normal range of frequency, duration, and strength, through effective compensatory mechanisms.[3,4]

There are two main methods for intrapartum FHR monitoring: continuous monitoring and intermittent auscultation (IA). Continuous monitoring is performed with cardiotocography (CTG), either external or internal, and is recommended in high-risk labor and birth. IA monitors the FHR at regular intervals during labor and is usually performed with a handheld Doppler device or a Pinard stethoscope. IA for FHR monitoring is recommended in low-risk births.[1,5,6] Many guidelines recommend that IA should be performed every 15 to 30 min, and for a duration of at least 1 min.[7–9] Using IA for FHR monitoring facilitates identification of the baseline FHR (baseline), and any fluctuations from the baseline, accelerations, and decelerations. IA is traditionally not regarded as appropriate for identifying type of decelerations or variability.[10]

Intrapartum FHR interpretation is complex and presents several challenges.[1] One challenge is the subjective human interpretation factor. Disagreement between raters interpreting FHR may lead to variations in interventions, as the interpretation guides clinical management decisions.[1,11] Reliability and agreement studies measure the consistency of assessments between different raters evaluating the same patients under similar conditions. The results can provide insights into the inherent measurement error and determine the validity of the test.[12] To the best of our knowledge, no studies have assessed reliability and agreement of intrapartum IA monitoring. A recent systematic review of reliability and agreement of intrapartum fetal monitoring[11] revealed large variations in human assessment of intrapartum CTG but found no studies assessing IA.

The aim of this study was to describe the interrater reliability and agreement in intrapartum IA monitoring using a handheld Doppler device. The primary aim was to measure the reliability and agreement of the baseline. The secondary aims were to determine the reliability and agreement of FHR increases and decreases from the baseline and classifications of the FHR sounds (normal or abnormal).

## 2 | METHODS

### 2.1 | Study and rater populations

We performed a pilot study to determine the sample size for the reliability study for the baseline, assessed as beats per minute (bpm). Eight midwives assessed 15 1-min FHR soundtracks once, from a form made with Nettskjema.[13] Nettskjema is a tool for designing and conducting online surveys and is operated by the University Information Technology Center (USIT) at the University of Oslo. A minimum of 154 fetal soundtracks and 16 raters were needed to achieve a target intraclass correlation coefficient (ICC) of 0.70 with a desired 95% confidence interval width of 0.10.[14] We also gained knowledge of the feasibility of the forms.

We used a convenient sampling method for the recruitment of raters, with help from research midwives and by posting information about our study at conferences and on social media. The inclusion criteria were midwives with experience using handheld Doppler for intrapartum IA. We aimed to include a sample of more than 16 raters to assess the FHR soundtracks, in case any of the raters withdrew after inclusion or did not complete the rating.

We used a convenient sampling method based on available and feasibly eligible women for the collection of FHR soundtracks. At several time points, from June 2019 through February 2023, we collected 215 soundtracks from 184 different women. We kept one soundtrack from each woman, excluded 47 of poor technical quality, and kept in total of 168 applicable soundtracks. Of these, 154 soundtracks were selected for the study. We aimed to include a sample of approximately 30% abnormal sounds to mimic the clinical field[15] and explore reliability and agreement on both normal and abnormal sounds.

The women were all admitted to Oslo University Hospital in Norway. The inclusion criteria for the women were singleton cephalic pregnancies at 37 to 42 weeks of gestation, from low-risk and high-risk women, admitted either for induction of labor or because of spontaneous onset of labor. A mixed population was chosen to facilitate recordings of soundtracks containing both normal and abnormal sounds.

To record the FHR soundtracks, we used a handheld Doppler device (Summit Doppler LifeDop 250 series) attached to a sound analyzer (Norsonic Nor140; Norsonic AS, Tranby, Norway).[16] The Doppler was placed on the woman's abdomen at a location where the fetal heart was heard most clearly. The FHR was recorded independent of the timing of maternal contractions.

The quality of the recordings was monitored by attaching headphones to the equipment. The recordings were started manually by the researcher and stopped

automatically by the sound analyzer. The audio files were stored in ".wav" format on an inserted SanDisk SD card (UHS-I card). In cases where the sound quality was unsatisfactory, such as unclear or excessively noisy recordings, a subsequent attempt was made. In 27 cases, we filmed the display of the Doppler, so the raters could see it while assessing the FHR. The displayed FHR is an average value calculated over a brief period and should help minimize variations caused by movements and other temporary factors and should provide a more reliable reading for the rater.[17]

## 2.2 | Rating process

The raters received written information about the aim of the study and a description of assessing IA during birth.[8,18] The sounds were stored anonymously on a computer, recoded, and transferred to Nettskjema.[13] We designed eight forms (Supporting Information S1) in Nettskjema[13] with 17–20 sounds in each form. The forms were distributed to the raters in random order.

The midwives were asked to assess the baseline as a single number. As the raters listened to decontextualized prerecorded soundtracks, the accurate detections of acceleration and deceleration were difficult to evaluate. As a substitute, they were asked to identify any increase (equivalent to acceleration) or decrease (equivalent to deceleration) from the baseline. A sound was defined as normal if the baseline was between 110 and 150 bpm, and the rhythm was regular with no decrease from the baseline. A sound was defined as abnormal if the baseline was above 150 or below 110, the rhythm was irregular, or with a decrease. A regular rhythm was further explained as a sound without major irregularities, and an irregular rhythm was explained as arrhythmic or uneven.[8,18] If the fetal sound was classified as abnormal, additional questions about the reason for the abnormality and intervention needed to be answered. The rating process started in February 2023 and ended in May 2023. The raters independently assessed the soundtracks one time and were blinded to each other's assessments. As in clinical practice, the sound quality varied to reflect what clinicians may encounter and to augment the generalizability of the findings.

## 2.3 | Statistical analysis

The primary end point, reliability of the baseline (bpm), was measured using an intraclass correlation coefficient (ICC) for agreement based on a two-way analysis of variance (ANOVA) model, denoted as ICC(A,1) with a 95% confidence interval (CI) determined by the moment approximation method.[19] ICC reflects both the degree of reliability of the measurement procedure and the agreement between measurements. Reliability refers to the ability of the Doppler to distinguish the FHR for women during labor. It shows how strongly repeated measurements made on the same participants resemble each other. ICC ranges from 0 to 1, and the higher the value, the stronger the reliability.[20] In this case, ICC also measured agreement in the sense that it took systematic differences between raters into account. The agreement between raters was further summarized by using the mean absolute distance between pairs of raters with 95% non-parametric percentile bootstrap confidence interval and the coverage probability,[21] giving the percentage of observations with an absolute difference between them when they were smaller than a predetermined value.

Secondary end points were FHR characteristics: baseline increase (yes, no, I don't know), baseline decrease (yes, no, I don't know), and classification (normal, abnormal, I don't know), which were all categorical measures. Agreements between the midwives were summarized using the proportion of agreement. The proportion of agreement gives the average percentage of items on which two raters agree. It varies between 0 and 1, and the higher the values, the stronger the agreement.[11] Reliability was assessed using Cohen's kappa coefficient. Cohen's kappa is sensitive to prevalence and is adjusted for chance agreement. Kappa values range from −1 to +1. Negative values indicate agreement lower than chance, a value of 0 indicates agreement no better than chance, and the higher values, the stronger the agreement.[11] We planned for subgroup analysis for fetal sounds with and without Doppler display, midwifery experience, reason for abnormality, type of abnormality, and type of intervention.

The results are reported following the Guidelines for Reporting Reliability and Agreement Studies (GRRAS).[11] Statistical analysis was performed in R (version 4.3.1 for Windows). The total deviation index was calculated using the Agreement package (version 0.8-1), and the ICC was calculated using the irr package (version 0.84.1). Missing values were not replaced.

We did not employ predetermined tables that assign labels like "poor," "moderate," or "good" for the interpretation of reliability and agreement values. We believe that acceptable agreement and reliability values are determined not solely by statistical criteria but also by clinical ones. Furthermore, the interpretation depends on the population studied and in which context the measurement will be used, and the predetermined tables do not take into account the uncertainty (confidence intervals).[11]

## 3 | RESULTS

In total, 24 midwives were recruited. Of these, four midwives withdrew before they started the assessment, and four withdrew during the study. This left a total of 16 raters independently assessing 154 FHR soundtracks one time each.

The 16 raters had from one to 40 years of midwifery experience, with an average of 14.4 years (SD: 14.2). Six raters (38%) had more than 10 years of experience. The midwives felt safe (69%) or very safe (31%) when using IA for fetal monitoring during birth. Most of the midwives replied that they count for 1 min (63%) or look at the Doppler display and count (56%) when they assess the FHR with IA in the clinical field (Table 1).

The assessments of the four midwives who withdrew after the inclusion were excluded from the analysis due to a considerable extent of missing values. Due to technical issues in the forms, one rater missed the assessment of seven soundtracks and two raters missed one soundtrack each. This led to 145 observations made by 16 raters for all outcomes.

**TABLE 1** Characteristics of the 16 raters.

| | *N* (%) |
|---|---|
| Midwife experience | |
| Years of experience >5 | 5 (31) |
| Years of experience 5–10 | 5 (31) |
| Years of experience 11–20 | 0 |
| Years of experience 21–30 | 3 (19) |
| Years of experience 31–40 | 3 (19) |
| Type of experience[a] (more than one answer possible) | |
| Obstetric unit, level 1 | 15 (94) |
| Obstetric unit, level 2 | 5 (31) |
| Midwifery unit, level 3 | 5 (31) |
| Home birth | 1 (6) |
| Antenatal maternity care | 3 (19) |
| Self-reported feeling of safety with IA | |
| Feeling unsafe | 0 |
| Feeling safe | 11 (69) |
| Feeling very safe | 5 (31) |
| Methods for IA in the clinical field (more than one answer possible) | |
| Counting 1 min | 10 (63) |
| Counting in frequencies | 2 (13) |
| Counting in frequencies and multiply[b] | 4 (25) |
| Looking at display and counting | 9 (56) |
| Looking at display and not counting | 6 (38) |
| Situation dependent | 6 (38) |

[a]Level 1: highly specialized units with advanced obstetric, pediatric, and anesthetic services. Level 2: birth units within smaller hospitals with obstetric and anesthetic services. Level 3: midwifery-led units caring for low-risk births.

[b]E.g., count for 15 s and multiply with 4.

Interrater reliability for the baseline is summarized in Table 2 and Figure 1. Overall, the ICC(A,1) was 0.74 (95% CI 0.69–0.78) (Table 2). When stratified by experience, midwives with 10 or fewer years of experience presented an ICC(A,1) of 0.75 (95% CI 0.70–0.80), while midwives with more than 10 years of experience presented an ICC(A,1) of 0.70 (95% CI 0.64–0.76). With an additional Doppler display, the ICC(A,1) was 0.91 (95% CI 0.85–0.95), while it was ICC(A,1) 0.70 (95% CI 0.64–0.76) without the display.

On average, an absolute difference of 7.9 bpm (95% CI 7.3–8.5) was observed between pairs of raters (Figure 2). The probability that the difference in FHR between two raters was fewer than 5 bpm was 35%, and fewer than 10 bpm was 63%.

Overall, an FHR increase was observed in 35% of the fetal sounds, and an FHR decrease in 14% and 65% of the fetal sounds were classified as normal (Table 2). The proportion of agreement (Po) for the FHR increase was 0.65 (95% CI 0.62–0.67) with a kappa of 0.23 (95% CI 0.18–0.27), and Po for the FHR decrease was 0.89 (95% CI 0.86–0.92) with a kappa of 0.54 (95% CI 0.43–0.64) (Table 2). The proportion of agreement for classification of the FHR sounds was 0.71 (95% CI 0.69–0.74) with a kappa of 0.37 (95% CI 0.31–0.43) (Table 2). Display and experience did not seem to influence the reliability and agreement measures for the secondary outcomes.

Overall, 91 unique numbers were used to assess the baseline of the 154 sounds. The midwives assessed the fetal sounds more often with numbers such as 140 and 130 than with numbers such as 141 and 131. For instance, the number 140 was used for assessment a total of 365 times, whereas 141 was used eight times, and 139 was used 18 times.

## 4 | DISCUSSION

### 4.1 | Main findings

The interrater reliability for the baseline (bpm) was ICC(A,1) 0.74 (95% CI 0.69–0.78). On average, an absolute difference of 7.9 bpm (95% CI 7.3–8.5 bpm) was observed between pairs of raters. Sixteen midwives evaluated the baseline of 154 fetal soundtracks one time each. The midwives had different midwifery experiences, but all were familiar with using the Doppler device and felt safe or very safe with IA for FHR monitoring.

**TABLE 2** Reliability and agreement measures of baseline FHR, FHR increases and decreases, and classification of FHR soundtracks.

| | *N* (%) | | | ICC(A,1)[a] (95% CI) |
|---|---|---|---|---|
| Baseline | | | | |
| All observers | 16 (100) | | | 0.74 (0.69–0.78) |
| <=10 years exp | 10 (62.5) | | | 0.75 (0.70–0.80) |
| >10 years exp | 6 (37.5) | | | 0.70 (0.64–0.76) |
| Doppler with display | 27 (17.5) | | | 0.91 (0.85–0.95) |
| Doppler without display | 127 (82.5) | | | 0.70 (0.64–0.76) |
| | | P (+)[b] | Po[c] (95% CI) | Kappa[d] (95% CI) |
| Increase[e] | | | | |
| All observers | 16 (100) | 0.35 | 0.65 (0.62–0.67) | 0.23 (0.18–0.27) |
| <=10 years exp | 10 (62.5) | 0.33 | 0.66 (0.63–0.68) | 0.22 (0.17–0.27) |
| >10 years exp | 6 (37.5) | 0.39 | 0.63 (0.59–0.66) | 0.23 (0.16–0.30) |
| Doppler with display | 27 (17.5) | 0.32 | 0.67 (0.60–0.73) | 0.24 (0.16–0.32) |
| Doppler without display | 127 (82.5) | 0.36 | 0.64 (0.62–0.67) | 0.22 (0.17–0.28) |
| Decrease[e] | | | | |
| All observers | 16 (100) | 0.14 | 0.89 (0.86–0.92) | 0.54 (0.43–0.64) |
| <=10 years exp | 10 (62.5) | 0.14 | 0.89 (0.86–0.92) | 0.54 (0.44–0.65) |
| >10 years exp | 6 (37.5) | 0.14 | 0.88 (0.85–0.92) | 0.51 (0.38–0.64) |
| Doppler with display | 27 (17.5) | 0.16 | 0.84 (0.76–0.91) | 0.39 (0.18–0.60) |
| Doppler without display | 127 (82.5) | 0.13 | 0.89 (0.87–0.93) | 0.58 (0.46–0.69) |
| Classification[e,f] | | | | |
| All observers | 16 (100) | 0.65 | 0.71 (0.69–0.74) | 0.37 (0.31–0.43) |
| <=10 years exp | 10 (62.5) | 0.68 | 0.75 (0.72–0.79) | 0.43 (0.35–0.51) |
| >10 years exp | 6 (37.5) | 0.61 | 0.65 (0.62–0.69) | 0.28 (0.21–0.36) |
| Doppler with display | 27 (17.5) | 0.68 | 0.75 (0.69–0.82) | 0.43 (0.27–0.59) |
| Doppler without display | 127 (82.5) | 0.65 | 0.70 (0.67–0.74) | 0.37 (0.29–0.43) |

Abbreviation: CI, confidence intervals.

[a]ICC(A,1), intraclass correlation coefficient for agreement based on a two-way ANOVA model.

[b]Mean proportion of positive cases over all raters.

[c]Proportion of agreement.

[d]Cohen's kappa coefficient.

[e]Grouping category no and I don't know together

[f]Classified into "normal," "abnormal," "I don't know."

## 4.2 | Strengths and limitations

This study was systematically and thoroughly planned, with an analysis plan established before the study. We performed a pilot on which we based sample size calculations for the main outcome. The raters were heterogenic in terms of clinical background and experience, and the FHR sounds varied in characteristics to reflect real clinical practice. This allowed for better external validity.

The sample size was based on assessment of the FHR baseline. The study could have benefited from involving more midwives and fetal sounds to decrease the width of the confidence intervals for the secondary outcomes. However, this was not possible due to time constraints. An important aspect of the baseline in a clinical context is the change in the baseline over time,[2] but this was not evaluated in this study. The raters were not provided with a specific definition of increase or decrease. As a result, this might influence varying interpretations of these terms.

## 4.3 | Interpretation

Interrater reliability and agreement studies provide information about consistency across subjective assessments between different raters. High levels of measurement indicate strong consistency between the raters. However, this does not guarantee the validity of their assessment, as

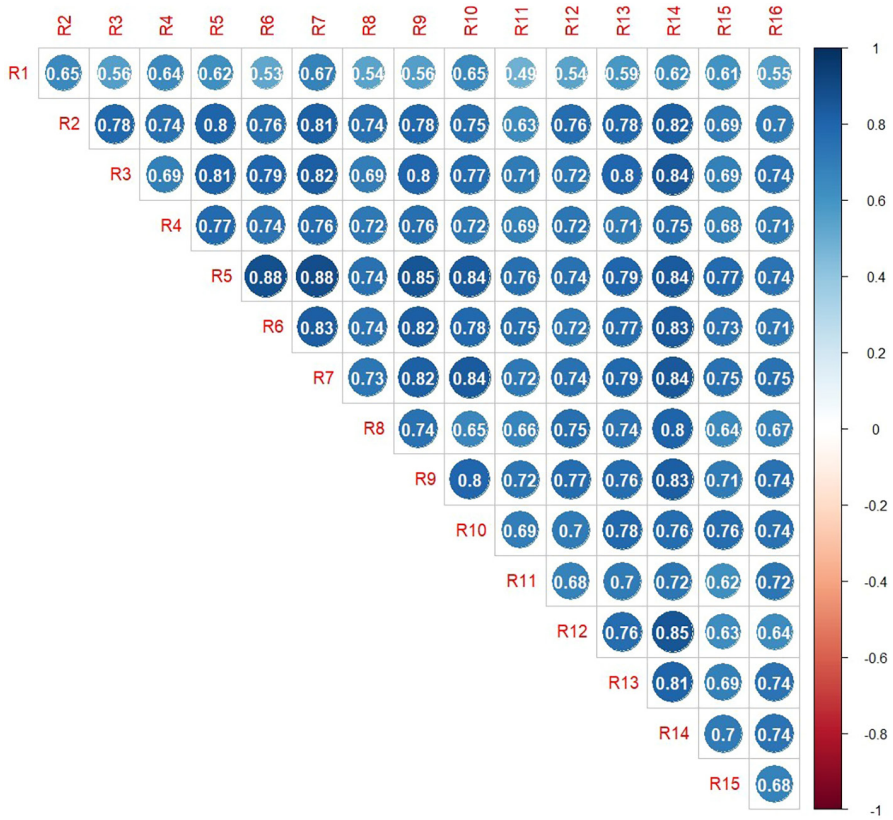


**FIGURE 1** Reliability of baseline FHR for pairs of raters (based on 16 raters and 145 completed observations). [Colour figure can be viewed at wileyonlinelibrary.com]

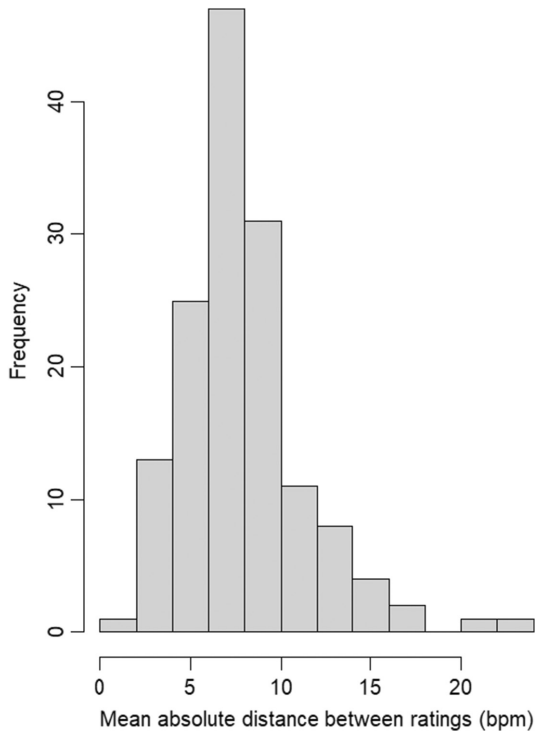


**FIGURE 2** Mean absolute distance (bpm) between ratings (agreement).

even unanimous agreement can be incorrect. Reliability and agreement studies serve as precursors to establishing validity, ensuring that the assessment is reliable before evaluating validity.[11,20]

Evaluating the baseline is important when assessing fetal well-being with IA,[10] as a normal baseline reflects the presence of normal oxygenation of the fetal myocardium.[2,4] If the baseline is near the borders for normal fetal heart rates (110–150 bpm), the mean absolute difference of 7.9 bpm (95% CI: 7.3–8.5 bpm) between pairs of raters found in this study is important information for clinicians, as it then may have implications on clinical care and interventions. While a single assessment slightly above (150 bpm) or below (110) the normal baseline definition might not lead to major intervention differences, variations over time could have a more significant impact and should warrant further assessment of the fetal well-being. A rise in the baseline can indicate fetal hypoxic stress due to catecholamine release,[3,4,10] often referred to as gradually evolving hypoxia. An increase in the baseline of more than 20 bpm from the start of labor[7] or an increase of >10% from the previously observed baseline[4,23] is proposed to be a sign of gradually evolving hypoxia. The mean absolute difference between raters in bpm was substantially lower than 20 bpm or >10%; 7.9 bpm is 5%–7% of total bpm given a normal FHR (110–150 bpm). This may indicate that the reliability and agreement found in this study are acceptable in detecting gradually evolving hypoxia. However, this remains to be verified in future studies, as this study did not examine changes in the baseline of fetal sounds.

Listening to prerecorded FHR sounds is different from actual clinical settings. In real life, midwives have the flexibility to listen longer, become familiar with the

baseline as they care for the woman during birth, reposition the Doppler for better clarity, and consider the clinical context. In this study, midwives heard sounds only once, relied on prerecorded quality, and lacked clinical context. Clinicians will argue that the well-being of a fetus can not only be assessed by listening to the FHR but must include other clinical findings.[4,22] The raters in our study commented that the absence of a clinical context in the sound assessment was challenging when trying to properly evaluate the sounds. This absence, however, specifically permits to assess the reliability and agreement of the baseline measurements, which was our primary goal.

We observed higher reliability for the baseline, ICC(A,1) 0.91 (95% CI 0.85–0.95), when the sounds were played with an additional Doppler display for 27 of the sounds. Sholapurkar[24,25] proposes that it is more informative to monitor the reading of the Doppler display than by just listening to the sound, but to the best of our knowledge, there have been no trials that assessed whether the display has an impact on neonatal outcomes. The higher reliability measures for the sounds with displays imply a higher consistency between the raters, but at the same time, just over half of the midwives in our study reported that they look at the display while counting. We did not have the statistical power to compare the results from the sounds with display versus those without display, and we did not find the same change in the measurements for the secondary outcomes.

In our study, the classification of the fetal sounds was measured to be Po 0.71 (95% CI 0.69–0.74), and kappa 0.37 (0.31–0.43). When abnormal sounds are identified, they should trigger further assessment, including more frequent auscultation and a comprehensive review of the clinical context.[7] The kappa value observed is not considered acceptable from a clinical perspective. It is important to note that classifying 1-min soundtracks can be challenging, especially when the context is absent. Our results identified difficulties when classifying sounds and are in line with other interrater reliability and agreement studies among healthcare professionals classifying intrapartum CTG,[12] lung sounds,[26] and heart sounds[27] revealing varying levels of agreement and reliability. Nonetheless, it is important to consider that measures across different studies are not directly comparable. A systematic review on reliability and agreement classifying intrapartum CTG traces[12] identified kappa values lower than expected by chance to nearly perfect agreement in 29 articles. A study on lung sound classification[26] involving 28 raters classifying sound recordings from 20 participants revealed kappa values ranging from 0.09 to 0.97 within subgroups of four raters. A study on heart sounds classification,[27] where 32 raters classified sounds from 200 participants, showed kappa ranges from 0.29 to 0.90.

The midwives often rounded the numbers to the nearest ten when evaluating the baseline, a phenomenon known as the approximate number system (ANS).[23] In addition, we observed that the midwives used different assessment methods, which, combined with ANS, might have contributed to variations between raters. Overall, we consider that the ICC and the absolute mean difference are acceptable values and believe that the result indicates that midwives can reasonably agree when assessing the baseline with a handheld Doppler. On the other hand, we recognize that a reliability measure of 0.74 and a 7.9-bpm difference could result in different clinical management decisions in certain situations. Assessing the reliability measures is thus not the same as testing the accuracy or assessing the outcomes, but it is essential before testing validity, as low reliability compromises validity.

## 5 | CONCLUSION

This study represents the first interrater reliability and agreement study of IA, to the best of our knowledge. Our results indicate an acceptable level of reliability when assessing the baseline using a handheld Doppler, an important feature of fetal well-being evaluation. In further research, it is worth investigating the impact of factors such as clinical context, numerical rounding, and the use of a Doppler display when counting and assessing FHR. In addition, examining the reliability and agreement concerning baseline trends over time could provide a more comprehensive understanding of assessing intrapartum fetal well-being with IA. Overall, our study contributes valuable insights into the reliability and agreement on IA for fetal monitoring. Enhancing reliability and agreement levels in IA may require standardizing the method and providing training for those who employ it.

### ACKNOWLEDGMENTS

We would like to acknowledge the midwives for their contributions to this study. Their commitment was pivotal in conducting this research. Appreciation is also extended to the participants who allowed access during their labor, a deeply personal event. We also recognize the efforts of the clinical midwives at Oslo University Hospital for facilitating the sound recording and the two midwives involved in the collection process.

### FUNDING INFORMATION

C. H. Engelhart received a PhD scholarship from The Norwegian Research Centre for Women's Health, Oslo University Hospital.

## CONFLICT OF INTEREST STATEMENT

There are no conflicts of interests to disclose in regard to this work.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

*Christina Hernandez Engelhart* https://orcid.org/0000-0002-5494-8783

## REFERENCES


1. Alfirevic Z, Gyte GML, Cuthbert A, Devane D. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst Rev*. 2017;2(2):CD006066.
2. Gracia-Perez-Bonfils A, Chandraharan E. Physiology of fetal heart rate control and types of intrapartum hypoxia. In: Chandraharan E, ed. *Handbook of CTG Interpretation: From Patterns to Physiology*. Cambridge University Press; 2017.
3. Jia Y-J, Ghi T, Pereira S, Gracia Perez-Bonfils A, Chandraharan E. Pathophysiological interpretation of fetal heart rate tracings in clinical practice. *Am J Obstet Gynecol*. 2023;228(6):622-644.
4. Griffiths K, Gupta N, Chandraharan E. Intrapartum fetal surveillance: a physiological approach. *Obstet Gynaecol Reprod Med*. 2022;32(8):179-187.
5. Ayres-de-Campos D, Spong CY, Chandraharan E. FIGO consensus guidelines on intrapartum fetal monitoring: cardiotocography. *Int J Gynecol Obstet*. 2015;131(1):13-24.
6. Lewis D, Downe S, FIGO Intrapartum Fetal Monitoring Expert Consensus Panel. FIGO consensus guidelines on intrapartum fetal monitoring: intermittent auscultation. *Int J Gynecol Obstet*. 2015;131(1):9-12.
7. National Institute for Clinical Excellence. *Fetal Monitoring in Labour*. National Institute for Health and Care Excellence; 2022 https://www.nice.org.uk/guidance/ng229
8. Kessler J, Blix E, Jettestad M, et al. Fosterovervåkning under fødsel, avnavling og syre-baseprøver fra navlesnor (fetal monitoring during birth, cord clamping and acid base samples). *Norwegian Assoc Gynecol Obstetr*. 2022. https://www.legeforeningen.no/foreningsledd/fagmed/norsk-gynekologisk-forening/veiledere/veileder-i-fodselshjelp/fosterovervakning-under-fodsel-avnavling-og-syre-baseprover-fra-navlesnor/
9. World Health Organization. *WHO Recommendations: Intrapartum Care for a Positive Childbirth Experience*. World Health Organization; 2018.
10. Lowe V, Archer A. Intermittent (intelligent) auscultation in the low-risk setting. In: Chandraharan E, ed. *Handbook of CTG Interpretation: From Patterns to Physiology*. Cambridge University Press; 2017.
11. Hernandez Engelhart C, Gundro Brurberg K, Aanstad KJ, et al. Reliability and agreement in intrapartum fetal heart rate monitoring interpretation: a systematic review. *Acta Obstet Gynecol Scand*. 2023;102(8):970-985.
12. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64(1):96-106.
13. University of Oslo. *Nettskjema*. University of Oslo; 2023 https://www.uio.no/english/services/it/adm-services/nettskjema/
14. Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. *Stat Med*. 2002;21(9):1331-1335.
15. Amer-Wåhlin I, Ingemarsson I, Marsal K, Herbst A. Fetal heart rate patterns and ECG ST segment changes preceding metabolic acidaemia at birth. *BJOG Int J Obstet Gynaecol*. 2005;112(2):160-165.
16. Norsonic. *Sound Analyser Nor140*. Norsonic; 2004 https://web2.norsonic.com/product_single/soundanalyser-nor140/
17. Medical M. Fetal heart rate display, Summit Doppler LifeDop 250 Auditory. In: Øian P, editor 2023.
18. American College of Nurse-Midwives. Intermittent auscultation for intrapartum fetal heart rate surveillance. *J Midwifery Womens Health*. 2015;60(5):626-632.
19. McGraw K, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1(1):30-46.
20. Koo TK, Li MY. A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163.
21. Lin LI-K. Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Stat Med*. 2000;19(2):255-270.
22. Engelhart CH, Nilsen ABV, Pay ASD, Maude R, Kaasen A, Blix E. Practice, skills and experience with the Pinard stethoscope for intrapartum Foetal monitoring: focus group interviews with Norwegian midwives. *Midwifery*. 2022;108:103288.
23. Odic D, Starr A. An Introduction to the approximate number system. *Child Dev Perspect*. 2018;12(4):223-229.
24. Chandraharan E. Fetal electrocardiograph (ST-Analyser or STAN): is it time for the requiem? *J Clin Med Surgery*. 2023;3:1111.
25. Sholapurkar SL. Intermittent auscultation (surveillance) of fetal heart rate in labor: a progressive evidence-backed approach with aim to improve methodology, reliability and safety. *J Matern Fetal Neonatal Med*. 2022;35(15):2942-2948.
26. Aviles-Solis JC, Vanbelle S, Halvorsen PA, et al. International perception of lung sounds: a comparison of classification across some European borders. *BMJ Open Respir Res*. 2017;4(1):e000250.
27. Andersen S, Davidsen AH, Schirmer H, Melbye H, Spigt M, Aviles-Solis JC. Interrater and intrarater agreement on heart murmurs. *Scand J Prim Health Care*. 2022;40(4):491-497.


## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Engelhart CH, Vanbelle S, Øian P, Pay ASD, Kaasen A, Blix E. How well can the fetal heart rate baseline be assessed by intrapartum intermittent auscultation? An interrater reliability and agreement study. *Birth*. 2024;51:835-842. doi:10.1111/birt.12858