Received: 12 January 2023 | Revised: 23 April 2023 | Accepted: 25 April 2023
DOI: 10.1111/aogs.14591

AOGS Acta Obstetricia et Gynecologica Scandinavica

SYSTEMATIC REVIEW

Reliability and agreement in intrapartum fetal heart rate monitoring interpretation: A systematic review

Christina Hernandez Engelhart[1,2] | Kjetil Gundro Brurberg[3] | Kristin Jerve Aanstad[2] | Aase Serine Devold Pay[2,4] | Anne Kaasen[2] | Ellen Blix[2] | Sophie Vanbelle[5]

[1]Norwegian Research Centre for Women's Health, Oslo University Hospital, Oslo, Norway
[2]Faculty of Health Sciences, Oslo Metropolitan University, Oslo, Norway
[3]Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway
[4]Department of Gynecology and Obstetrics, Bærum Hospital, Vestre Viken Hospital Trust, Gjettum, Norway
[5]Department of Methodology and Statistics, Maastricht University, Maastricht, Netherlands

Correspondence
Christina Hernandez Engelhart, Oslo universitetssykehus, Rikshospitalet, Nasjonalt senter for kvinnehelseforskning, Postboks 4950 Nydalen, 0424 Oslo, Norway.
Email: cheng@oslomet.no

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

Abstract

Introduction: Fetal heart rate (FHR) monitoring is routine in intrapartum care worldwide and one of the most common obstetrical procedures. Intrapartum FHR monitoring helps assess fetal wellbeing and interpretation of the FHR help form decisions for clinical management and intervention. It relies on the observers' subjective assessments, with variation in interpretations leading to variations in intrapartum care. The purpose of this systematic review was to summarize and evaluate extant inter- and intrarater reliability research on the human interpretation of intrapartum FHR monitoring.

Material and Methods: We searched for the terms "fetal heart rate monitoring," "interpretation agreement" and related concepts on Embase, Medline, Maternity and Infant Care Database and CINAHL. The last search was made on January 31, 2022. The protocol for the study was prospectively registered in PROSPERO (CRD42021260937). Studies that assess inter- and intrarater reliability and agreement of health professionals' intrapartum FHR monitoring were included and studies including other assessment of fetal wellbeing excluded. We extracted data in reviewer pairs using quality appraisal tool for studies of diagnostic reliability (QAREL) forms. The data retrieved from the studies are presented as narrative synthesis and in additional tables.

Results: Forty-nine articles concerning continuous FHR monitoring were included in the study. For interrater reliability and agreement, in total 577 raters assessed 6315 CTG tracings. There was considerable heterogeneity in quality and measures across the included articles. We found higher reliability and agreement for the basic FHR features than for overall classification and higher agreement for intrarater reliability and agreement than for their interrater counterparts.

Conclusions: There is great variation in reliability and agreement measures for continuous intrapartum FHR monitoring, implying that intrapartum CTG should be used with caution for clinical decision making given its questionable reliability. We found

Abbreviations: CTG, cardiotocography; FHR, fetal heart rate; IA, intermittent auscultation; ICC, intraclass correlation coefficients; Pa, proportion of agreement; Ps, proportion of specific agreement; QAREL, quality appraisal tool for studies of diagnostic reliability; κ, kappa statistics.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2023 The Authors. *Acta Obstetricia et Gynecologica Scandinavica* published by John Wiley & Sons Ltd on behalf of Nordic Federation of Societies of Obstetrics and Gynecology (NFOG).

few high-quality studies and noted methodological concerns in the studies. We recommend a more standardized approach to future reliability studies on FHR monitoring.

KEYWORDS
agreement, fetal heart rate, fetal monitoring, interrater and intrarater, observer variation, reliability, reproducibility of results, systematic review

1 | INTRODUCTION

Intrapartum fetal heart rate (FHR) monitoring is a critical component of the assessment of fetal wellbeing, aimed at identifying inadequately oxygenated fetuses to facilitate timely intervention to prevent fetal damage. There are two main methods for assessing FHR: intermittent auscultation (IA) and continuous electronic monitoring via cardiotocography (CTG).[1] IA was the main method for fetal monitoring during labor until CTG was introduced into clinical practice in the late 1960s.[1] Pinard stethoscopes and hand-held Doppler devices are the most common instruments used for IA.[2] CTG is used alone or in conjunction with other methods, such as fetal scalp blood sampling and ST segment analysis of the fetal electrocardiogram. CTG is one of the most common obstetrical procedures and, along with IA, is routine in intrapartum care worldwide.[1,3,4]

FHR assessment is a clinical observation that depends on observers' subjective skills and clinical guidelines for interpretation.[1] When assessing FHR through IA, the observer evaluates heart rate baseline, rhythm, and the presence and absence of accelerations and decelerations.[5] With CTG, the observer evaluates the basic FHR features of baseline, variability, accelerations, and decelerations, as well as maternal uterine contractions. Based on these features, observers derive an overall CTG classification that determines if the CTG tracing is normal or abnormal.[6] As interpretations of IA and CTG help form decisions for clinical management and intervention, any variation in them will lead to variations in intrapartum care. Consequences of these variations may result in excessive, inappropriate, or lack of appropriate interventions.[1]

Health professionals' interpretations of a clinical test include measurement error, which can be quantified using reliability and agreement. Reliability refers to the degree to which a measurement procedure can distinguish between patients, despite measurement error.[7] Agreement refers to the closeness of repeated measurements of the same patients made under similar conditions.[7] Interrater reliability and agreement studies involve multiple observers (raters) who evaluate the same patients in similar conditions, while intrarater reliability and agreement studies involve repeated measurements made by a single observer of the same patients.[7]

To our knowledge, there is no systematic review that assesses observer variability in human interpretations of intrapartum FHR monitoring. A consolidated look at existing research might reveal where interpretation needs to be improved. The aim of this systematic review was thus to summarize and evaluate extant inter- and intrarater reliability research on the human interpretation of intrapartum FHR monitoring.

Key message

There is a diversity of reliability and agreement studies of intrapartum fetal monitoring interpretation. The studies are heterogenetic, with wide variations of reliability and agreement.

2 | MATERIAL AND METHODS

This study was conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) 2020 checklist[8] and Meta-analysis Of Observational Studies in Epidemiology (MOOSE) guidelines.[9] The protocol was prospectively registered in PROSPERO (CRD42021260937).

2.1 | Search strategy and data sources

We systematically searched the electronic databases Embase (Ovid), Medline (Ovid MEDLINE ALL), Maternity and Infant Care Database (Ovid), and CINAHL (Ebsco) for relevant literature. A senior medical librarian at the University of Oslo performed the searches in August 2021 and again on January 31, 2022, to search for new published articles. Controlled vocabulary (MeSH, Emtree terms) and free-text search terms of the two concepts "fetal heart rate monitoring" and "interpretation agreement" were combined with the Boolean operator AND. The search was not restricted by language. The search strategy was designed to identify all studies on reliability, agreement, and validity/accuracy as these terms are often used incorrectly. The search strategy is detailed in Table S1.

2.2 | Eligibility criteria

Two reviewers (CHE and EB) independently screened the titles and abstracts derived from the database searches using Rayyan (Qatar Computing Research Institute), a web and mobile app for systematic reviews.[10] Reviewer pairs (CHE & EB, CHE & AK, CHE & ASDP, CHE & KJA, CHE & SV) assessed articles in full for inclusion. Disagreements were resolved via consensus through discussions, with a third reviewer invited if needed.

We included all quantitative studies that assess the inter- and intrarater reliability of intrapartum FHR monitoring, irrespective of

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

study design, study setting, type of observers, or reported statistical measure. The studies had to be available in languages understood by the reviewers (Dutch, English, French, German, Spanish, or a Scandinavian language). Studies including assessments of fetal wellbeing other than FHR, duplicates, unpublished articles, gray literature, abstracts, and non-scientific material were excluded.

2.3 | Quality assessment and data extraction

We used the quality appraisal tool for studies of diagnostic reliability (QAREL) for the quality appraisal and data extraction.[11,12] The reviewers tested and agreed upon criteria for the interpretation of the items in the form. The quality appraisal tool consists of 11 items, and a higher score indicates higher quality (Table S2a,b). A data extraction form was used in combination with the QAREL forms[12] and featured 23 items, all related to the 11 items in the QAREL, that covered design and setting, type of reliability, population characteristics, observers, types of tests, statistics used, and the appropriateness of the statistics reported. The reviewer pairs independently assessed the studies' methodological quality and extracted data from each eligible article. Disagreements were discussed until consensus was reached. We contacted the study authors for clarification where necessary.

2.4 | Data analysis

We reported all statistical measures used in the selected reliability and agreement studies. For studies of intrapartum FHR monitoring via CTG, we examined results for FHR baseline, variability, acceleration, deceleration, and overall tracing classification. For those which concern IA, we assessed FHR baseline, acceleration, deceleration, rhythm, and overall heart rate classification. We planned to conduct subgroup analyses based on FHR assessment method (IA or CTG), profession, experience, training, and guidelines used, and to perform meta-analyses of studies that were sufficiently similar from a clinical and statistical standpoint.

3 | RESULTS

3.1 | Study selection

The electronic literature search resulted in 2671 articles, and a manual search of reference lists yielded two additional articles. The screening procedure is described in Figure 1, and the PRISMA flow diagram and our reasons for full-text exclusions are given in Table S3. After screening, and assessing 151 articles in full text, we included 49 articles about

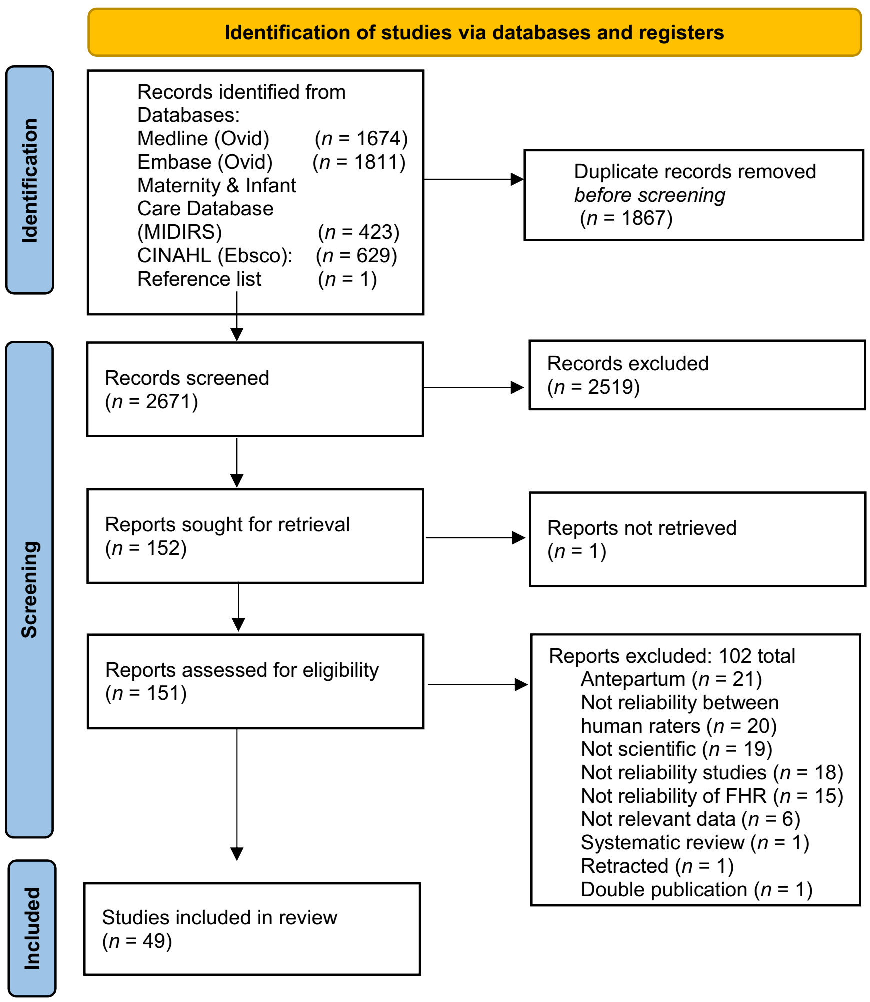

FIGURE 1 PRISMA 2020 flow diagram. *Source*: Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the reliability and agreement of CTG, but did not identify any eligible articles that assess reliability or agreement for IA.

3.2 | Study characteristics

The studies were conducted in Europe, North America, Asia, and Oceania in various clinical settings. The study populations varied in terms of risk, gestational age, stage of labor, and neonatal outcomes. The observers were midwives, nurses, physicians, medical students, and "clinical experts" (e.g., had expertise in fetal monitoring). In total, 577 raters assessed 6315 CTG tracings (interrater reliability and agreement), and 123 raters (one study with an unknown number of raters) assessed 1170 CTG tracings twice (intrarater reliability and agreement). We identified 17 different guidelines used to interpret CTG tracings, of which four had two editions from different years[13–19] and some had similar content.[19–22] These guidelines are described in Table S4. Table 1 outlines the detailed characteristics of the 49 included articles.

3.3 | Statistical reliability and agreement measures

Statistical measures to assess reliability and agreement varied across the articles (Table 1), with proportion of agreement (Pa), kappa statistics (κ), proportion of specific agreement (Ps), and intraclass correlation coefficients (ICC) most often reported. Pa is an absolute agreement measure reporting the proportion of cases in which raters agree.[24] K corrects that proportion for agreements expected by chance and ranges from −1 to +1, with +1 indicating perfect chance agreement, 0 being equivalent to chance, and negative estimates representing under chance agreement. Weighted κ further attributes different weights to disagreements according to the magnitude of the disagreement.[24] Ps gives the probability that two raters assign a participant to a category given that at least one of the raters assigned the participant to that category.[25] Lastly, ICC tells how strongly repeated measurements made on the same subjects resemble each other. Its minimum and maximum values are 0 and 1, respectively.[26]

Several of the included articles used prearranged tables to interpret κ values (e.g., Landis & Koch).[27] We decided not to use these tables as the interpretation of κ depends on study context and the population studied, and one table cannot account for the variety across heterogeneous studies. In addition, the prearranged tables do not include confidence interval interpretations.[7]

3.4 | Reliability and agreement in intrapartum fetal monitoring

The data retrieved from the included studies are presented in Table 1, Tables S5–S10 and as narrative synthesis. The groups used in the narrative synthesis (CTG tracing classification, FHR baseline, FHR variability, FHR decelerations, FHR accelerations), are based on clinical relevance and how the included studies grouped their results. As the included studies were excessively heterogenetic, in terms of population, methods and statistics used, we did not perform any meta-analysis.

The interrater reliability and agreement assessment of the CTG tracings were grouped per classification ($n=33$ studies; Table S5), baseline ($n=17$; Table S6), variability ($n=21$; Table S7), decelerations ($n=21$; Table S8), and accelerations ($n=15$; Table S9). Interrater reliability and agreement were described using κ, Pa, Ps, and ICC.

Guidelines for the classification of CTG tracings were based on different tier-classification systems, with tiers equivalent to the classification categories used for interpretation (e.g., normal, suspicious, and pathological). The most frequently used guideline was International Federation of Gynecology and Obstetrics (FIGO) 2015 ($n=15$; Table S5).[14]

Of the 33 articles that presented classification measures, 29 presented measures for overall tracing classification. The κ coefficients and Pa varied considerably, with κ values ranging from lower than excepted by chance to almost perfect reliability and agreement (Table S5).

Thirteen articles assessed agreement on a specific category in the tier-classification system. In most studies, higher κ values and Pa were more frequent for normal CTG classifications than for abnormal classifications (Table S5).

Nine articles[15,28,31–37] reported an association between type of guideline used for interpretation and interrater reliability. For κ coefficients, we did not find any consistency across articles between κ magnitude and the type of guideline used, but in two studies reporting Pa, the American College of Obstetricians and Gynecologists (ACOG) was associated with the highest agreement level[28,36] (Table S5). We present extracted κ coefficient measures with FIGO guidelines in Table 2.

In general, reliability and agreement were higher for FHR baseline compared to classification, variability, and decelerations. We also found a higher Pa for normal baseline classifications as compared to abnormal classifications[30,36,38] (Table S6).

The reported reliability for variability differed considerably. For agreement, we found a higher Pa for normal variability classifications than for abnormal classifications,[30,36,38–40] except for one measurement in one study where reduced variability showed the highest agreement when κ was used (Table S7).[39]

Several studies grouped reliability and agreement according to type of deceleration. In general, the variable deceleration had the lowest κ coefficient and the variable late deceleration had the lowest Pa. Prolonged decelerations showed the largest κ and Pa values (Table S8).

Reliability and agreement for FHR accelerations varied, but mostly yielded large κ values, ICC, and Pa (Table S9).

We found 14 articles that assessed intrarater reliability. The CTG tracing assessments were grouped per classification ($n=11$ studies), baseline ($n=2$), variability ($n=4$), decelerations ($n=2$), and accelerations ($n=2$; Table S10). For variability, baseline, acceleration, and

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Characteristics of included studies ($n = 49$).

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Amer-Wahlin etal.[55](2005), Sweden	Compare the rates of abnormal CTG and ST patterns in acidemia cases and controls and assess the reproducibility of ST and CTG assessments	Interrater $N = 142$; relative high risk, indication for scalp monitoring, singleton cephalic, >36 GA	Interrater $R = 2$; obstetricians	Interrater; classification	FIGO 1987	Proportion of agreement (Pa)	9/11
Ayres-de-Campos etal.[56](1999), Portugal	Evaluate interobserver agreement in expert interpretation of CTG tracings following the FIGO guideline, and the subsequent clinical decision	Interrater $N = 17$; high risk, indication for scalp monitoring, singleton cephalic	Interrater $R = 3$; experts	Interrater; classification, decelerations	FIGO 1987	Cohen's kappa, weighted kappa (linear weights), proportion of specific agreement (Ps) (Chamberlain)	8/11
Ayres-de-Campos etal.[48](2004), Portugal	Evaluate the reproducibility of FHR baseline estimation according to an objective and detailed definition presented in the article, by comparison with the FIGO guidelines' definition	Interrater $N = 150$; high risk, unselected population, singleton, 35–42 GA	Interrater $R = 9$; clinician specialist in obstetrics and gynecology	Interrater; baseline	FIGO 1987	Cohen's kappa, Ps (Chamberlain), ICC (unclear form)	8/11
Beaulieu etal.[58](1982), Canada	Determine the degree to which intrapartum CTGs are consistently assessed	Interrater $N = 150$, intrarater $N = 150$; normal and abnormal tracings	Interrater $R = 5$, intrarater $R = 5$; obstetricians	Interrater; classification Intrarater; classification	Own clinical criteria	Pa (interrater between the 5 reviewers), Pa (intrarater for each reviewer)	9/10 10/11
Bernardes etal.[38](1997), Portugal	Evaluate interobserver agreement in visual analysis of CTG event	Interrater $N = 17$; high risk, 3. trimester	Interrater $R = 3$; obstetricians, expertise in fetal monitoring	Interrater; baseline, variability, decelerations, accelerations	FIGO 1987	Cohen's kappa, Ps (Chamberlain)[a]	7/11
Bhatia et al.[32] (2017), UK	Compare FIGO 2015, NICE 2007, and NICE 2014 guidelines	Interrater $N = 10$; high risk, mixed outcome	Interrater $R = 21$; midwives, obstetricians	Interrater; classification	FIGO 2015, NICE 2007, NICE 2014	Cohen's kappa, Pa	4/11
Blackwell et al.[39] (2011), USA	Assess the interobserver and intraobserver reliability of the NICHD 3-tier FHR classification system	Interrater $N = 40$, intrarater $N = 20$; different UA pH, ≥37 GA	Interrater $R = 3$, intrarater $R = 3$; practitioners	Interrater; classification, variability, decelerations, accelerations. Intrarater; classification	NICHD 2008	Cohen's kappa, kappa for each category (unclear), Pa	8/11 9/11
Blix etal.[59](2003), Norway	Examine the agreement in assessment of the labour admission test between midwives and obstetricians in the clinical setting and two experts in the non-clinical setting	Interrater $N = 845$; mixed population, at admission, >28 GA	Interrater $R = 5$; midwives, obstetricians	Interrater; classification	Ingemarsson and Ingemarsson 1987	Weighted kappa (linear weights), Ps (Chamberlain)	8/11
Blix and Øian[59] (2005), Norway	Examine interobserver agreements when the labor admission tests were assessed by midwives and obstetricians who had received training in interpreting CTG traces	Interrater $N = 549$; mixed population, at admission, >28 GA	Interrater $R = 6$; midwives, obstetricians	Interrater; classification	Ingemarsson 1986, Sundstrøm 2000	Weighted kappa (linear weights), Ps (Chamberlain)	8/11

(Continues)

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Burrus et al.[16] (1994), USA	Analyze associations between intrapartum FHR tracings and short- and long-term outcome in neonates delivered between 24 and 26 weeks of gestation	Interrater $N=41$; high risk, last hour of recordings, 24–26 GA	Interrater $R=2$; board-certified MFM specialist	Interrater; classification, baseline, variability, decelerations	Parer 1983, Kublie et al. 1969	Cohen's kappa	8/11
Buscicchio et al.[60] (2012), Italy	Assess reproducibility and clinical relevance of current guidelines on FHR interpretation in labor	Interrater $N=100$; first 100 labors of the month	Interrater $R=2$; medical doctors	Interrater; classification	NICE 2007	Cohen's kappa	4/11
Chauhan et al.[61] (2008), USA	Determine interobserver variability in the classification of FHR tracing with periodic deceleration as being reassuring or non-reassuring and in the ability to predict emergency caesarean delivery or umbilical arterial pH <7.00	Interrater $N=100$; non-reassuring FHR, singleton cephalic, ≥34 GA	Interrater $R=5$; clinician doctors	Interrater; classification, baseline, accelerations	ACOG 2005	Weighted kappa (unclear weights)	8/11
Chen etal.[62](2014), Taiwan	Compare a novel computerized analysis program with visual CTG interpretation results	Interrater $N=62$; no known medical problems or congenital anomalies, at admission, ≥37 GA	Interrater $R=8$; obstetricians	Interrater; classification, baseline, variability, decelerations, accelerations	NICHD 2008	Cohen's kappa, ICC (unclear form)	8/11
Devane and Lalor[63] (2005), Ireland	Examine intra- and interobserver agreement in midwives' visual interpretations of intrapartum CTGs	Interrater $N=3$, intrarater $N=3$; reproduced tracings	Interrater $R=28$, intrarater $R=28$; midwives	Interrater; classification, baseline, variability, decelerations. Intrarater; classification	FIGO 1987	Kappa (unclear type for interrater), Cohen's kappa (intrarater)	5/11 5/11
Devoe et al.[46] (2000), USA	Compare the visual analyses of FHR tracings by observers according to recent NICHD interpretative guidelines both with each other and with those of a computerized FHR analysis and alerting system	Interrater $N=50$; indications for continuous electronic FHR monitoring, singleton, active phase of labor, ≥32 GA	Interrater $R=4$; RN, CNM, OB resident physician, physician MFM faculty member	Interrater; baseline, decelerations, accelerations	NICHD 1997	Pa (between pairs)	7/11
Donker et al.[65] (1993), The Netherlands	Assess interobserver variation in the assessment of FHR recordings	Interrater $N=10$; cross-section of obstetric situation	Interrater $R=21$, obstetricians	Interrater; classification, variability	Unclear	Fleiss kappa, average pairwise Pa	4/11
Ekengård etal.[15](2021), Sweden	Compare the templates, SWE-09, FIGO-15 and SWE-17, regarding sensitivity and specificity in identifying acidosis during the first stage of labor	Interrater $N=292$; cesarean section, pH <7.10, singleton, 1. stage of labor, ≥34 GA	Interrater $R=3$; midwives, physicians	Interrater; classification	FIGO 2015, SWE-09, SWE-17	Free-marginal kappa, Pa	9/11
Epstein et al.[52] (2013), USA	Evaluate the interobserver reliability of FHR pattern definition and interpretation assessed by physicians at various levels of training using standard NICHD definitions and standard principles of interpretation	Interrater $N=32$; singleton, 5 h preceding delivery, at term	Interrater $R=5$; medical students, residents, junior specialists, senior medicine specialists	Interrater; classification, baseline, variability, decelerations, accelerations	NICHD 2008	Free marginal kappa, Spearman correlation, Pa	7/11

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Epstein et al.[44] (2016), USA	Compare the interobserver reliability among groups of obstetrics care providers with the use of the contemporary Hon-Quilligan method of FHR interpretation with the Caldeyro-Barcia method with the use of an online interactive testing tool	Interrater $N=40$; full spectrum of fetal outcomes from normal to poor, singleton, ≥36 GA	Interrater $R=15$; obstetric care providers	Interrater; baseline, variability, decelerations, accelerations	Caldeyro-Barcia 1966, Hon-Quilligan Caldeyro (ACOG 2009)	Adjusted Gwet-Kappa, Cohen's kappa, Pa (between pairs), averaged pairwise Pa	4/11
Escarena et al.[65] (1979), USA	Design a scheme to score visual interpretations of the FHR tracing for variability content, determine the extent of agreement among different physicians using a proposed scoring system and compare the visually scored tracings and the computed variability estimates	Interrater $N=12$; collected from a wide field of FHR variability types, late labor, 37–42 GA	Interrater $R=9$; experts	Interrater; variability	Unclear	Fleiss kappa	5/11
Farquhar et al.[23] (2020), New Zealand	Determine whether experienced clinicians could detect abnormal CTG readings taken during the penultimate hour before delivery in babies diagnosed with moderate to severe NE and recommend an appropriate action plan	Intrarater $N=30$; tracings from neonates with NE and without NE or birth hypoxia, singleton and multiple, 1-h penultimate birth, ≥37 GA	Intrarater $R=10$; midwives, obstetricians	Intrarater; classification	RANZCOG 2014	Cohen's kappa, Pa	7/11
Flu et al.[66] (1990), The Netherlands	Assess inter- and intraobserver variations for intrapartum FHR variables	Interrater $N=100$, intrarater $N=100$; high risk, at term	Interrater $R=2$, intrarater $R=2$; obstetricians	Interrater; baseline, variability. Intrarater; baseline, variability decelerations, accelerations	Caldeyro-Barcia et al 1966, Hon andQuilligan1968 (modificated)	Cohen's kappa	4/11 5/11
Garabedian etal.[28](2017), France	Assess the interobserver reliability of those four FHR classifications	Interrater $N=100$; singleton, last 60 min prior to pushing, >37 GA	Interrater $R=4$; obstetricians	Interrater; classification	ACOG 1995, CNGOF 2013, FIGO 2015, NICHD 2008	Krippendorff's alpha, Cohen's kappa (between pairs), Cohen's kappa (weighted average), weighted kappa (unclear weights)	8/11
Ghi et al.[32] (2016), Italy	Assess retrospectively the accuracy of the RCOG and the Piquard CTG classification systems in identifying a group of fetuses delivered in the second stage of labor with metabolic acidemia at birth.	Interrater $N=246$; fetuses delivered with metabolic acidemia, singleton, 2. stage of labor, ≥37 GA	Interrater $R=2$; senior and trainee operators	Interrater; classification	RCOG (NICE 2014), PIQUARD 1988	Cohen's kappa, Pa	8/11

(Continues)

AOGS Acta Obstetricia et Gynecologica Scandinavica

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Govindappagari et al.[43] (2016), USA	Determine if the online EFM course required by our medical malpractice insurance company, completed by all labor and delivery staff, translated into improved standardization of FHR monitoring interpretation	Interrater $N=701$; singleton, admission and before delivery	Interrater $R=$unknown; nurses, physicians (variety of teams)	Interrater; variability, decelerations, accelerations	NICHD 2008	Cohen's kappa, Pa	4/11
Gyllencreutz etal.[49](2017), Sweden	Examine whether the combination of web-based CTG education and on-site CTG training could lead to a better agreement in CTG interpretation than web-based education alone	Interrater $N=106$, intrarater $N=106$; indication for FBS, singleton cephalic, ≥34 GA	Interrater $R=6$, intrarater $R=6$; obstetricians	Interrater; baseline, variability, decelerations, accelerations. Intrarater; baseline, variability, decelerations, accelerations	Own clinical criterial	Cohen's kappa, weighted kappa (unclear weights) (intrarater), Fleiss kappa (interrater)	8/11 8/11
Gyllencreytz etal.[68](2018), Sweden	Evaluate the reliability of the computerized algorithm in the identification and characterization of FHR decelerations and to compare the reliability with manual/ visual assessment	Interrater $N=4$; indication for FBS, singleton cephalic, ≥34 GA	Interrater $R=2$; obstetricians	Interrater; decelerations	Not applicable	Bland–Altman analysis, ICC (two-way mixed model)	5/11
Hall et al.[68] (2012), USA	Develop a prototype electronic ruler to aid clinician assessment of FHR variability on electronic monitors and assess the performance of this prototype ruler with expert clinicians	Interrater $N=30$, intrarater $N=10$, at term	Interrater $R=6$, intrarater $R=6$; obstetricians/ gynecologists	Interrater; variability Intrarater; variability	NICHD 2008	Weighted kappa (unclear weights)	7/11 7/11
Hayashi et al.[34] (2012), Japan	Assess the reproducibility and clinical usefulness of 5-tier classification proposed by the Perinatology Committee of the Japan Society of Obstetrics and Gynecology (JSOG)	Interrater $N=107$, intrarater $N=107$; singleton, active labor, >34 GA	Interrater $R=2$, intrarater $R=2$; obstetricians	Interrater; classification Intrarater; classification	JSOG 2010, subjective	Weighted kappa (quadratic weights)	8/11 8/11
Hruban et al.[70] (2015), Czech Republic	Evaluate obstetricians' inter- and intraobserver agreement on intrapartum CTG recordings and to examine obstetricians' evaluations with respect to umbilical artery pH and base deficit	Interrater $N=552$, intrarater $N=82$; singleton, >37 GA	Interrater $R=9$, intrarater $R=9$; experienced obstetricians	Interrater; classification Intrarater; classification	FIGO 1987	Fleiss kappa, median Pa between pairs	4/11 5/11
Kundu etal.[42](2017), Germany	Analyze the intra- and interobserver variability of obstetricians and midwives with different professional experience by classifying the CTGs in the last 60 min before delivery	Interrater $N=300$, intrarater $N=300$; singleton cephalic, >37 GA	Interrater $R=7$, Intrarater $R=7$; midwives, obstetricians	Interrater; classification Intrarater; classification	FIGO 2015	Intrarater and interrater variance	7/11 7/11

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Lawson etal.[70](2000), Canada	Examine the interobserver reliability among Ontario obstetricians in the interpretation of intrapartum CTGs	Interrater $N = 12$; reassuring and non-reassuring tracings	Interrater $R = 74$; obstetricians	Interrater; decelerations	Unclear	Fleiss kappa, Maximum Pa	7/11
Lemoine etal.[72](2016), France	Assessing inter- and intraobserver agreement in the reading of FHR between two different paper speeds (1 and 2 cm/min) using FIGO classification	Interrater $N = 60$; high risk, >36 GA	Interrater $R = 6$; midwives, obstetricians	Interrater; classification	FIGO 1987	Cohen's kappa, Weighted kappa (linear weights), Ps (Chamberlain)	8/11
Marti Gamboa et al.[29] (2017), Spain	Compare the new 3-tier system with the 5-tier system, and to determine which of both systems have the greater ability to detect neonatal acidemia and which has better interobserver agreement identifying those fetuses at risk of neurological damage	Interrater $N = 202$; neonatal acidemia and not neonatal acidemia, singleton, last 30 min of monitored labor, at term	Interrater $R = 2$; obstetricians	Interrater; classification	FIGO 2015, Parer 2007	Kappa (unclear type)	8/11
Nielsen etal.[73](1987), Denmark	Estimate the magnitude of the intra- and interobserver variability among obstetricians actively engaged in the EFM and to estimate the accuracy of the interpretation of a specifies period of the CTG	Interrater $N = 50$, intrarater $N = 50$; mixed CTG patterns and newborn outcomes, last 30 min of 1. stage, 36–43 GA	Interrater $R = 4$, intrarater $R = 4$; obstetricians	Interrater; classification Intrarater; classification	Unclear	Fisher's one-tailed test, McNemar's test	3/11 3/11
Ojala etal.[74](2008), Finland	Evaluate the interobserver variability in the assessment of STAN recordings	Interrater $N = 200$; non-selected women, singleton cephalic, >36 GA	Interrater $R = 3$; experienced consultant	Interrater; classification	STAN (NICE 1997)	Cohen's kappa, weighted kappa (linear weights), Pa	8/11
Palomaki etal.[38](2006), Finland	Examine interobserver variation in visual interpretation of intrapartum CTG readings	Interrater $N = 22$; (mainly) mixed population	Interrater $R = 31$; obstetricians	Interrater; baseline, variability, decelerations	Unclear	Pairwise Ps (Chamberlain)	7/11
Peleg et al.[75] (2016), Israel	Determine if the chart speed affects electronic fetal monitor interpretation	Interrater $N = 19$; high risk, at term	Interrater $R = 14$; physicians	Interrater; classification, variability, decelerations, accelerations	ACOG 2009	Free marginal kappa, Pa	7/11
Rei etal.[30](2016), Portugal	Evaluate interobserver agreement in interpretation of CTG tracings using the new 2015 FIGO guidelines on intrapartum fetal monitoring	Interrater $N = 151$; indication for continuous CTG, singleton, last 60 min before delivery, ≥36 GA	Interrater $R = 6$; clinicians	Interrater; classification, baseline, accelerations variability, decelerations	FIGO 2015	Light's kappa, Pa, Ps (Dice)	8/11
Reif etal.[42](2016), Austria, France, Slovenia, Belgium, Portugal	Evaluate if knowledge of neonatal umbilical artery pH value affects the retrospective analysis of CTG tracings according to the NICE guidelines, and whether it influences subsequent clinical management recommendations	Interrater $N = 42$; uneventful antepartum, spontaneous labor, singleton cephalic, last 60 min before delivery, ≥ 37 GA	Interrater $R = 123$; midwives, obstetricians	Interrater; classification	NICE 2008	Cohen's kappa	8/11

(Continues)

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Rhose et al.[47] (2014), The Netherlands	Quantify inter- and intraobserver agreement in the classification of intrapartum CTG patterns prior to FBS, according to the FIGO/STAN guidelines, and management based on this classification	Interrater $N=79$, intrarater $N=79$; high risk, non-reassuring CTG, singleton cephalic, 1. stage of labor, 60-min prior to FBS, ≥37 GA	Interrater $R=9$, intrarater $R=9$; midwives, obstetricians	Interrater; classification Intrarater; classification	STAN 2007 (FIGO 1987)	Weighted kappa (unclear weights), Ps (Chamberlain)	4/11 5/11
Sabiani etal.[35](2015), France	Evaluate the intra- and interobserver agreement among obstetric experts in court regarding the retrospective review of abnormal FHR tracings and obstetrical management of patients with abnormal FHR during labor	Interrater $N=30$, intrarater $N=30$; abnormal FHR singleton cephalic, >37 GA	Interrater $R=22$, intrarater $R=22$; obstetrical experts in court	Interrater; classification	FIGO 1987, CNGOF 2008	Kappa (multiple raters-unclear type)	9/11 9/11
Santo etal.[35](2017), Portugal (UK)	Compare interobserver agreement, reliability and accuracy of CTG analysis, when performed according to the FIGO, ACOG and NICE guidelines	Interrater $N=151$; indication for CTG, singleton cephalic, last 60 min of tracings obtained before delivery, ≥37 GA	Interrater $R=27$; clinicians	Interrater; classification, baseline, variability, decelerations, accelerations	ACOG 2009, FIGO 1987, NICE 2007	Light's kappa, Ps (Dice), Pa	7/11
Schiermeier etal.[41](2011), Germany	Evaluate and compare different judges with different work experience and different training with non-invasive computer analyzing software using the FIGO classification	Interrater $N=12$	Interrater $R=33$; midwives, obstetricians	Interrater; baseline, variability, decelerations, accelerations	FIGO 1987	Ps (Chamberlain)[a]	8/11
Taylor etal.[76](2000), Ireland	Estimate the reliability of components of the FHR trace, and the validity of a computerized algorithm as regards these components	Interrater $N=24$; intrapartum or undergoing induction of labor	Interrater $R=7$; senior obstetric staff	Interrater; baseline, variability, decelerations, accelerations	FIGO 1987	Cohen's kappa, ICC (unclear form)	6/11
Vejux etal.[37](2017), France	Compare intrapartum CTG analysis in case of first cesarean section for non-reassuring CTG according to international guidelines	Interrater $N=100$; cesarean indication for non-reassuring CTG, singleton cephalic, ≥34 GA	Interrater $R=4$; obstetricians	Interrater; classification	FIGO 1987, CNGOF 2008	Weighted kappa (unclear weights)	8/11
Westerhuis et al.[77] (2009), The Netherlands	Quantify inter- and intraobserver agreement on classification of the intrapartum CTG and decision to intervene following STAN guidelines	Interrater $N=73$, intrarater $N=73$, high risk, >36 GA	Interrater $R=6$, intrarater $R=6$; medical doctors	Interrater; classification Intrarater; classification	STAN (FIGO 1987)	Cohen's kappa, Ps (Dice)	8/11 9/11
Wolfberg et al.[78] (2008), USA	Develop a computerized algorithm to quantify FHR variability and compare it to perinatologists' interpretation of FHR variability	Interrater $N=30$, intrarater $N=30$; scalp for clinical indication, singleton, 35–41 GA	Interrater $R=4$, intrarater $R=4$; perinatologists	Interrater; variability Intrarater; variability	NICHD 1997	Weighted kappa (unclear weights) Pearson's correlation, ICC (unclear form)	7/11 8/11

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 (Continued)

Author (year), country	Objective	Subjects	Raters	Reliability	Guideline	Reliability and agreement measures	QAREL Quality appraisal
Zamora del Pozo et al.[31] (2021), Spain	Evaluate interobserver agreement and the capacity to predict neonatal acidemia of the patterns categorized as pathological from the updated cardiotocographic guidelines FIGO in 2015, ACOG in 2010, NICE in 2017, and the new guideline based on fetal physiology by Chandraharan in 2018	Interrater $N=150$; mixed pH ranges, singleton cephalic, at term	Interrater $R=3$; expert reviewers	Interrater; classification, baseline, variability, decelerations, accelerations	ACOG 2009, FIGO 2015, NICE 2017, Chandraharan	Fleiss kappa	8/11

Note: The characteristics of the included studies only presents characteristics used in the synthesis.

Abbreviations: ACOG, The American College of Obstetricians and Gynecologists; CNGOG, Le Collège National des Gynécologues et Obstétriciens Français; CNM, Certified Nurse Midwife; CTG, cardiotocogram; EFM, electronic fetal monitoring; FBS, fetal blood sampling; FHR, fetal heart rate; FIGO, International Federation of Gynecology and Obstetrics; GA, gestational age; ICC, intraclass correlation coefficient; JSOG, Perinatology Committee of the Japan Society of Obstetrics and Gynecology; MFM, Maternal Fetal Medicine; NE, neonatal encephalopathy; NICE, National Institute for Health and Care Excellence; NICHD, National Institute of Child Health and Human Development; OB, obstetric; pH, potential of hydrogen; RANZCOG, Royal Australian and New Zealand College of Obstetricians and Gynecologists; RCOG, Royal College of Obstetricians and Gynecologists; RN, registered nurse; STAN, ST-analysis; SWE, National template Sweden; UA, umbilical artery.

[a]The papers cite Gant, understood as Proportion of specific agreement (Chamberlain). Chamberlain can only be computed for specific categories, but the results in the study are computed globally.

deceleration, the κ coefficients were high. For overall classification of the CTG tracings, κ and Pa varied but were mostly high.

Five articles assessed reliability in relation to rater experience[30,36,41–43] and six in relation to rater profession (Tables S5–S10).[41,42,44–47] In general, across the articles we did not find any clear association between rater experience or profession and reliability. In turn, we found three articles assessing reliability of FHR baseline, variability, and accelerations in relation to pre- and post-training sessions,[44,48,49] where reliability and agreement were generally higher after training sessions.

3.5 | Methodological quality

The results of the quality assessment of the included studies using QAREL are described in Table S2. Two of the 11 items in the QAREL (items 4 and 9) were not relevant to the interobserver reliability articles, and one (item 9) was not relevant to the intraobserver reliability articles. The quality scores ranged from 3 to 9 for inter- and 3 to 10 for intraobserver reliability. We found that variations in quality scores were mainly due to insufficient reporting (Table S2).

4 | DISCUSSION

We reviewed 49 articles that examine inter- and intrarater reliability and agreement for intrapartum FHR monitoring interpretation. No studies assessing IA monitoring met our inclusion criteria. The studies were of different methodological quality, with low to high quality scores according to the QAREL checklist.

We found considerable heterogeneity in the study populations and reliability reported in the articles in term of patient population and statistical methods used. Due to the high heterogeneity we decided not to present results from meta-analyses. Materials and methods were generally reported inadequately, particularly regarding subject population. Many of the studies did not report confidence intervals. The κ coefficient, Pa, Ps, and ICC were the most frequently reported measures of reliability and agreement.

The four basic FHR features and overall CTG tracing classifications were interpreted using 17 different clinical guidelines. For interrater reliability, we found that the studies reported higher reliability (κ and ICC) and agreement (Pa and Ps) for basic FHR features than for overall CTG tracing classifications. Most of the interrater reliability studies showed higher agreement (Pa and Ps) in normal tracing classifications, baseline, and variability than in abnormal classifications. We also found generally higher intra- than interrater reliability. We did not find any clear association in the studies between reliability and rater experience or profession, but higher reliability was achieved after training sessions.

The studies included used subjective FHR pattern assessment as a measurement instrument. This assessment is commonly interpreted according to clinical guidelines in which FHR patterns and uterine contractions (CTG) are evaluated. Guidelines are usually

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Results of interobserver variation in classifying intrapartum cardiotocograph with FIGO guideline.

Authors and year	Guideline / Additional grouping (population)	Reliability estimates: Kappa (95% CI) Overall, global	Normal	Suspicious	Pathological
Ayres-de-Campos et al.[57] (1999)	FIGO 1987	0.31[a,b] (0.11–0.51)			
Bhatia et al.[32] (2017)	FIGO 2015	0.38[a]			
Devane and Lalor[64] (2005)	FIGO 1987	0.69	0.54	0.77	0.75
Ekengård et al.[15] (2021)	FIGO 2015				
	Acidemic	0.47[c] (0.32–0.62)			
	Non acidemic	0.91[c] (0.87–0.96)			
Garabedian et al.[28] (2017)	FIGO 2015	0.59[d] (0.49–0.67)			
Hruban et al.[70] (2015)	FIGO 1987	0.255[e] (0.253–0.258)			
Lemoine et al.[72] (2016)	FIGO 1987				
	1 cm/min	0.42			
	1 cm/min weighted kappa	0.54 (0.44–0.64)			
	1 cm/min complete	0.22			
	2 cm/min	0.39			
	2 cm/min weighted kappa	0.51 (0.40–0.63)			
	2 cm/min complete	0.28			
Marti Gamboa et al.[29] (2017)[f]	FIGO 2015	0.466	0.568	0.288	0.538
Rei et al.[30] (2016)	FIGO 2015	0.39[g] (0.33–0.45)			
Sabiani et al.[35] (2015)[f]	FIGO 1987				
	Last 60 min before birth				
	All cases	0.13[h] (0.10–0.16)			
	Adverse outcome	0.13[h] (0.10–0.16)			
	Last 30 min before birth				
	All cases	0.12[h] (0.07–0.16)			
	Adverse outcome	0.12[h] (0.08–0.16)			
Santo et al.[36] (2017)	FIGO 1987	0.37[g] (0.31–0.43)			
Vejux et al.[37] (2017)	FIGO 1987	0.331[i] (0.27–0.39)			
Zamora del Pozo et al.[31] (2021)	FIGO 2015	0.35[e] (0.28–0.41)	0.46 (0.36–0.55)	0.29 (0.20–0.39)	0.29 (0.20–0.38)

Abbreviations: FIGO, International Federation of Gynecology and Obstetrics.
[a]Cohen's Kappa.
[b]Weighted kappa (linear weights).
[c]Free-marginal kappa.
[d]Krippendorf alpha.
[e]Fleiss kappa.
[f]Not all data are extracted.
[g]k-Light's kappa.
[h]Kappa (multiple raters'-unclear type).
[i]Weighted kappa (unclear weights).

developed through expert consensus and used by diverse health-care professionals.[4,50] We identified 17 guidelines in this systematic review but did not find any clear association between the type of guideline used for interpretation and level of reliability. Intrarater reliability and agreement levels were higher than their interrater counterparts, meaning that the same rater was more consistent when interpreting the same tracing twice than different raters who interpreted the same tracing. This might reflect the subjectivity of interpretations, where one rater will likely interpret and adopt the same guideline each time, whereas different raters might have different understandings of the same guideline.[50]

When exploring disagreements between raters, we did not find any clear association between rater experience and profession, but agreement could be improved through training sessions. Kelly

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

et al.[51] reviewed the impact of intrapartum CTG training, finding that it has a favorable effect on participant knowledge and skills and that it improves interobserver reliability compared to no training. The certainty of these two pieces of evidence was considered low and very low, respectively, but training is recognized to ensure the appropriate use of CTG.[14]

We found that the raters reached higher reliability and agreement for basic FHR features than for overall CTG tracing classifications. Notably, to classify a CTG, all four FHR features and uterine contractions need to be evaluated.[14] Our results indicate that it is easier to assess and interpret one basic FHR feature than it is to make a more complex interpretation of multiple features. In fact, the variations in the measurements reported for overall CTG classification, points to a weakness of intrapartum FHR monitoring. CTG classifications is an important aspect of intrapartum care, as it is used as basis for intrapartum intervention decisions. This is important to emphasize, particularly considering the widespread use of CTG. Variations in the interpretation of FHR monitoring, will probably affect the consistency of intrapartum care. This is further complicated by the fact that it appeared easier for the raters to agree on normality than abnormality, as most of the studies reported a higher Ps when the tracings, baseline, and variability were classified as normal. In fact, the real strength of intrapartum FHR monitoring might lie in its prediction of the absence of fetal metabolic acidemia.[1,52] FHR patterns are also sensitive indicators with limited specificity when predicting fetal hypoxia[1,14] and FHR interpretations might be complicated by their pattern complexity. In a recent study, Johnson et al.[53] points to this fundamental weakness in the use of electronic FHR monitoring. The authors question if further interpretation improvement will enhance the usefulness of continuous FHR monitoring and significantly alter clinical outcomes. The wide biological variability in the fetus's ability to tolerate intrapartum hypoxic stress also leads to an unpredictable and highly variable individual threshold for injury outcomes and may have less to do with clinicians' inadequate pattern interpretation.

It is correspondingly important to emphasize that FHR patterns should be interpreted in conjunction with maternal, fetal, and external factors for a comprehensive understanding of fetal wellbeing and appropriate management in the real clinical world.[14,20] The raters in the included studies mostly interpreted CTG tracings outside a clinical context; they were thus taken out of a potentially stressful environment and had the opportunity to discuss the situation with colleagues. This is a common means of performing reliability studies, as it allows for exact reliability measurements.[32] However, it might also affect external validity and generalizability, as reliability in a real-life context might differ.

To our knowledge, this is the first systematic review to assess reliability and agreement in intrapartum FHR monitoring. A strength of this review is its comprehensive systematic literature search, which placed no restriction on type of study, language, or publication date. Reviewer pairs assessed all the included articles, and we have all presented a thorough review of our findings.

However, this study still has limitations. We did not identify articles on IA assessing reliability and agreement, which limited our original intentions of reviewing articles about both CTG and IA. There was considerable heterogeneity across the included articles, which meant that we were not able to perform the intended meta-analysis. In addition, several of the articles had scarce data regarding setting, methods, and study population. We found great variability in the reported measures of reliability as well, and in the included studies' quality scores. We did not grade quality of evidence as the data found in this systematic review do not fit any existing grading framework. Thus, our results should be interpreted carefully.

We noted that several of the included studies did not meet the Guidelines for Reporting Reliability and Agreement Studies (GRRAS), which is a frequently recognized challenge within these types of studies.[7] In particular, the type of statistical measure used was sometimes not sufficiently detailed (e.g., the type of weights used for κ, whether the Ps or Pa was computed for specific categories) or not appropriate (e.g., Pearson's correlation coefficient was used as a reliability measure). The subject population was also often inadequately reported.

The included studies further reported a variety of reliability and agreement measures. First, some studies inappropriately used Pearson's correlation, McNemar's test, and Fisher's one-tailed test as reliability or agreement measures. Second, the concepts of agreement and reliability were sometimes misused interchangeably, though they are two different concepts.[7] Among the statistical agreement measures presented in the included studies, only κ and quadratic weighted κ can be interpreted as reliability measures, as defined in Lord and Novick's[53,54] classical test theory. Further, reliability is a measure specific to the population studied and can only be generalized to populations with similar characteristics.[7] In addition, interrater reliability studies often included only two raters, thus calling into question the generalizability of their conclusions to other raters. In sum, the included articles' clinical and statistical heterogeneity, and the wide variations in reliability measures without estimating uncertainty, made interpretation and syntheses of the results difficult.

5 | CONCLUSION

There is currently a lack of high-quality studies that evaluate both inter- and intraobserver variation when assessing intrapartum FHR monitoring via CTG. Among the existing articles, we found reliability and agreement measures to vary from almost perfect to worse than chance. Additionally, there was considerable variation in the CTG classification measures. This implies that intrapartum CTG should be used with caution for clinical decision making given its questionable reliability.

Furthermore, we also found methodological concerns in the included studies, and recommend a more standardized approach to future reliability studies on FHR with more thorough reporting of methodological details, especially regarding subject populations. Improved reporting will enable stronger comparisons across studies,

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

potentially leading to more accurate and reliable examination of monitoring methods.

AUTHOR CONTRIBUTIONS

CHE, AK, ASDP, KJA and EB conceptualized and designed the study. CHE and EB developed the search strategy and screened titles and abstracts. CHE, AK, ASDP, KJA, EB and SV extracted the data from the included articles. CHE wrote the original draft. All authors analyzed and interpreted the data, and critically reviewed and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank Hilde Iren Flaatten, the senior medical librarian at the University of Oslo, who helped us to develop the search strategy and performed the searches.

FUNDING INFORMATION

C. H. Engelhart received a PhD scholarship from the Norwegian Research Center for Women's Health at Oslo University Hospital, but the Research Center was not involved in the completion of this study.

CONFLICT OF INTEREST STATEMENT

The authors confirm that there are no conflicts of interest.

ORCID

Christina Hernandez Engelhart https://orcid.org/0000-0002-5494-8783
Anne Kaasen https://orcid.org/0000-0001-9291-0105
Ellen Blix https://orcid.org/0000-0001-7971-4580

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:Cd006066.
2. Martis R, Emilia O, Nurdiati DS, Brown J. Intermittent auscultation (IA) of fetal heart rate in labour for fetal well-being. *Cochrane Database Syst Rev*. 2017;2:CD008680.
3. Garabedian C, De Jonckheere J, Butruille L, Deruelle P, Storme L, Houfflin-Debarge V. Understanding fetal physiology and second line monitoring during labor. *J Gynecol Obstet Hum Reprod*. 2017;46:113-117.
4. Ayres-de-Campos D, Arulkumaran S. FIGO consensus guidelines on intrapartum fetal monitoring: introduction. *Int J Gynecol Obstet*. 2015;131:3-4.
5. Blix E, Maude R, Hals E, et al. Intermittent auscultation fetal monitoring during labour: a systematic scoping review to identify methods, effects, and accuracy. *PLoS ONE*. 2019;14:e0219573.
6. Ayres-de-Campos D, Arulkumaran S, FIGO Intrapartum Fetal Monitoring Expert Consensus Panel. FIGO consensus guidelines on intrapartum fetal monitoring: physiology of fetal oxygenation and the main goals of intrapartum fetal monitoring. *Int J Gynaecol Obstet*. 2015;131:5-8.
7. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol*. 2011;64:96-106.
8. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
9. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of Observational Studies in Epidemiology (MOOSE) group. *JAMA*. 2000;283:2008-2012.
10. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210.
11. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Med Res Methodol*. 2013;13:111.
12. Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). *J Clin Epidemiol*. 2010;63:854-861.
13. Rooth G, Huch A, Huch R. Guidelines for the use of fetal monitoring. FIGO News. *Int J Gynaecol Obstet*. 1987;25:159-167.
14. Ayres-de-Campos D, Spong CY, Chandraharan E. FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. *Int J Gynecol Obstet*. 2015;131:13-24.
15. Ekengård F, Cardell M, Herbst A. Impaired validity of the new FIGO and Swedish CTG classification templates to identify fetal acidosis in the first stage of labor. *J Matern Fetal Neonatal Med*. 2021;35:4853-4860.
16. Parer JT. *Handbook of Fetal Heart Rate Monitoring*. Philadelphia Saunders; 1983.
17. Burrus DR, O'Shea TM Jr, Veille JC, Mueller-Heubach E. The predictive value of intrapartum fetal heart rate abnormalities in the extremely premature infant. *Am J Obstet Gynecol*. 1994;171:1128-1132.
18. Parer JT, Ikeda T. A framework for standardized management of intrapartum fetal heart rate patterns. *Am J Obstet Gynecol*. 2007;197:26.e1-26.e6.
19. National Collaborating Centre for Women's and Children's Health (UK). *National Institute for Health and Clinical Excellence: uidance*. RCOG Press; 2007 (NICE clinical guidelines, No. 55.).
20. National Collaborating Centre for Women's and Children's Health (UK). *Intrapartum care for healthy women and babies*. National Institute for Health and Care Excellence (NICE); 2014. Available from: https://www.nice.org.uk/guidance/cg190/chapter/Recommendations#monitoring-during-labour
21. Macones GA, Hankins GDV, Spong CY, Hauth J, Moore T. The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. *J Obstet Gynecol Neonatal Nurs*. 2008;37:510-515.
22. ACOG. Practice Bulletin No. 106: intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. *Obstet Gynecol*. 2009;114:192-202.
23. Farquhar CM, Armstrong S, Masson V, Thompson JMD, Sadler L. Clinician identification of birth asphyxia using intrapartum cardiotocography among neonates with and without encephalopathy in New Zealand. *JAMA Netw Open*. 2020;3:e1921363.
24. Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall; 1991.
25. Grant JM. The fetal heart rate trace is normal, isn't it? Observer agreement of categorical assessments. *Lancet*. 1991;337:215-218.
26. Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. *Biometrics*. 1980;36:19-25.
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.
28. Garabedian C, Butruille L, Drumez E, et al. Inter-observer reliability of 4 fetal heart rate classifications. *J Gynecol Obstet Hum Reprod*. 2017;46:131-135.
29. Marti Gamboa S, Gimenez OR, Mancho JP, Moros ML, Sada JR, Mateo SC. Diagnostic accuracy of the FIGO and the 5-tier fetal

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

heart rate classification systems in the detection of neonatal acidemia. *Am J Perinatol*. 2017;34:508-514.
30. Rei M, Tavares S, Pinto P, et al. Interobserver agreement in CTG interpretation using the 2015 FIGO guidelines for intrapartum fetal monitoring. *Eur J Obstet Gynecol Reprod Biol*. 2016;205:27-31.
31. Zamora Del Pozo C, Chóliz Ezquerro M, Mejía I, et al. Diagnostic capacity and interobserver variability in FIGO, ACOG, NICE and Chandraharan cardiotocographic guidelines to predict neonatal acidemia. *J Matern Fetal Neonatal Med*. 2021;35:8498-8506.
32. Bhatia M, Mahtani KR, Nunan D, Reddy A. A cross-sectional comparison of three guidelines for intrapartum cardiotocography. *Int J Gynaecol Obstet*. 2017;138:89-93.
33. Ghi T, Morganelli G, Bellussi F, et al. Cardiotocographic findings in the second stage of labor among fetuses delivered with acidemia: a comparison of two classification systems. *Eur J Obstet Gynecol Reprod*. 2016;203:297-302.
34. Hayashi M, Nakai A, Sekiguchi A, Takeshita T. Fetal heart rate classification proposed by the perinatology committee of the Japan Society of Obstetrics and Gynecology: reproducibility and clinical usefulness. *J Nippon Med Sch*. 2012;79:60-68.
35. Sabiani L, Le Dû R, Loundou A, et al. Intra- and interobserver agreement among obstetric experts in court regarding the review of abnormal fetal heart rate tracings and obstetrical management. *Am J Obstet Gynecol*. 2015;213:856.e1-856.e8.
36. Santo S, Ayres-de-Campos D, Costa-Santos C, et al. Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines. *Acta Obstet Gynecol Scand*. 2017;96:166-175.
37. Vejux N, Ledu R, D'Ercole C, Piechon L, Loundou A, Bretelle F. Guideline choice for CTG analysis influences first caesarean decision. *J Matern Fetal Neonatal Med*. 2017;30:1816-1819.
38. Palomäki O, Luukkaala T, Luoto R, Tuimala R. Intrapartum cardiotocography—the dilemma of interpretational variation. *J Perinat Med*. 2006;34:298-302.
39. Bernardes J, Costa-Pereira A, Ayres-de-Campos D, van Geijn HP, Pereira-Leite L. Evaluation of interobserver agreement of cardiotocograms. *Int J Gynaecol Obstet*. 1997;57:33-37.
40. Blackwell SC, Grobman WA, Antoniewicz L, Hutchinson M, Gyamfi BC. Interobserver and intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system. *Am J Obstet Gynecol*. 2011;205:378.e1-378.e5.
41. Schiermeier S, Westhof G, Leven A, Hatzmann H, Reinhard J. Intra- and interobserver variability of intrapartum cardiotocography: a multicenter study comparing the FIGO classification with computer analysis software. *Gynecol Obstet Invest*. 2011;72:169-173.
42. Reif P, Schott S, Boyon C, et al. Does knowledge of fetal outcome influence the interpretation of intrapartum cardiotocography and subsequent clinical management? A multicentre European study. *BJOG*. 2016;123:2208-2217.
43. Kundu S, Kuehnle E, Schippert C, von Ehr J, Hillemanns P, Staboulidou I. Estimation of neonatal outcome artery pH value according to CTG interpretation of the last 60 min before delivery: a retrospective study. Can the outcome pH value be predicted? *Arch Gynecol Obstet*. 2017;296:897-905.
44. Govindappagari S, Zaghi S, Zannat F, et al. Improving Interprofessional consistency in electronic fetal heart rate interpretation. *Am J Perinatol*. 2016;33:808-813.
45. Epstein AJ, Iriye BK, Hancock L, et al. Web-based comparison of historical vs contemporary methods of fetal heart rate interpretation. *Am J Obstet Gynecol*. 2016;215:488.e1-488.e5.
46. Devoe L, Golde S, Kilman Y, Morton D, Shea K, Waller J. A comparison of visual analyses of intrapartum fetal heart rate tracings according to the new national institute of child health and human development guidelines with computer analyses by an automated fetal heart rate monitoring system. *Am J Obstet Gynecol*. 2000;183:361-366.
47. Rhöse S, Heinis AM, Vandenbussche F, van Drongelen J, van Dillen J. Inter- and intra-observer agreement of non-reassuring cardiotocography analysis and subsequent clinical management. *Acta Obstet Gynecol Scand*. 2014;93:596-602.
48. Ayres-de-Campos D, Bernardes J, Marsal K, et al. Can the reproducibility of fetal heart rate baseline estimation be improved? *Eur J Obstet Gynecol Reprod*. 2004;112:49-54.
49. Gyllencreutz E, Hulthén Varli I, Lindqvist PG, Holzmann M. Reliability in cardiotocography interpretation—impact of extended on-site education in addition to web-based learning: an observational study. *Acta Obstet Gynecol Scand*. 2017;96:496-502.
50. Santo S, Ayres-de-Campos D. Human factors affecting the interpretation of fetal heart rate tracings: an update. *Curr Opin Obstet Gynecol*. 2012;24:84-88.
51. Kelly S, Redmond P, King S, et al. Training in the use of intrapartum electronic fetal monitoring with cardiotocography: systematic review and meta-analysis. *BJOG*. 2021;128:1408-1419.
52. Epstein AJ, Twogood S, Lee RH, Opper N, Beavis A, Miller DA. Interobserver reliability of fetal heart rate pattern interpretation using NICHD definitions. *Am J Perinatol*. 2013;30:463-468.
53. Johnson GJ, Salmanian B, Denning SG, Belfort MA, Sundgren NC, Clark SL. Relationship between umbilical cord gas values and neonatal outcomes: implications for electronic fetal heart rate monitoring. *OBGYN*. 2021;138:366-373.
54. Lord F, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley; 1968.
55. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33:613-619.
56. Amer-Wahlin 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.
57. Ayres-de-Campos D, Bernardes J, Costa-Pereira A, Pereira-Leite L. Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision. *BJOG*. 1999;106(12):1307-1310.
58. Beaulieu MD, Fabia J, Leduc B, et al. The reproducibility of intrapartum cardiotocogram assessments. *Can Med Assoc J*. 1982;127(3):214-216.
59. Blix E, Sviggum O, Koss KS, Øian P. Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts. *BJOG: Int J Obstet Gynaecol*. 2003;110(1):1-5.
60. Blix E, Øian P. Interobserver agreements in assessing 549 labor admission tests after a standardized training program. *Acta Obstet Gynecol Scand*. 2005;84(11):1087-1092.
61. Buscicchio G, Gentilucci L, Martorana R, Martino C, Tranquilli AL. How to read fetal heart rate tracings in labor: a comparison between ACOG and NICE guidelines. *J Matern Fetal Neonatal Med*. 2012;25(12):2797-2798.
62. Chauhan SP, Klauser CK, Woodring TC, Sanderson M, Magann EF, Morrison JC. Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability. *Am J Obstet Gynecol*. 2008;199(6):623.e1-623.e5.
63. Chen CY, Yu C, Chang CC, Lin CW. Comparison of a novel computerized analysis program and visual interpretation of cardiotocography. *PLoS ONE*. 2014;9(12):e112296 (Electronic resource).
64. Devane D, Lalor J. Midwives' visual interpretation of intrapartum cardiotocographs: intra- and inter-observer agreement. *J Adv Nurs*. 2005;52(2):133-141.
65. Donker DK, van Geijn HP, Hasman A. Interobserver variation in the assessment of fetal heart rate recordings. *Eur J Obstet Gynecol Reprod*. 1993;52(1):21-28.
66. Escarena L, McKinney RD, Depp R. Fetal baseline heart rate variability estimation. I. Comparison of clinical and stochastic quantification techniques. *Am J Obstet Gynecol*. 1979;135(5):615-621.
67. Flu PK, Bohnen AM, Wallenburg HCS. Intrapartum fetal heart rate patterns. I. Classification, quantification and observer variation. *J Matern Fetal Neonatal Med*. 1990;3(4):209-215.

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

68. Gyllencreutz E, Lu K, Lindecrantz K, et al. Validation of a computerized algorithm to quantify fetal heart rate deceleration area. *Acta Obstet Gynecol Scand*. 2018;97(9):1137-1147.
69. Hall LM, Hannon DJ, Leisk GG, Wolfberg AJ, House MD. Measurement of fetal heart rate variability on an electronic monitor using a prototype electronic ruler. *Am J Perinatol*. 2012;29(6):409-413.
70. Hruban L, Spilka J, Chudáček V, et al. Agreement on intrapartum cardiotocogram recordings between expert obstetricians. *J Eval Clin Pract*. 2015;21(4):694-702.
71. Lawson H, Fairley S, Morris K. Inter-observer agreement in cardiotocogram interpretation: how reliable is Ontario? *J SOGC*. 2000;22(5):366-373.
72. Lemoine H, Ehlinger V, Groussolles M, Arnaud C, Vayssiere C. Does the paper speed in fetal heart monitoring during labour influence the variability in the interpretation by professionals? *J Gynecol Obst Bio R*. 2016;45(8):827-834.
73. Nielsen PV, Stigsby B, Nickelsen C, Nim J. Intra- and inter-observer variability in the assessment of intrapartum cardiotocograms. *Acta Obstet Gynecol Scand*. 1987;66(5):421-424.
74. Ojala K, Mäkikallio K, Haapsamo M, Ijäs H, Tekay A. Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies. *Acta Obstet Gynecol Scand*. 2008;87(5):536-540.
75. Peleg D, Ram R, Warsof SL, et al. The effect of chart speed on fetal monitor interpretation. *J Matern Fetal Neonatal Med*. 2016;29(10):1577-1580.
76. Taylor GM, Mires GJ, Abel EW, et al. The development and validation of an algorithm for real-time computerized fetal heart rate monitoring in labour. *BJOG: Int J Obstet Gynaecol*. 2000;107(9):1130-1137.
77. Westerhuis M, van Horen E, Kwee A, van der Tweel I, Visser G, Moons K. Inter- and intra-observer agreement of intrapartum ST analysis of the fetal electrocardiogram in women monitored by STAN. *BJOG Int J Obstet Gynaecol*. 2009;116(4):545-551.
78. Wolfberg AJ, Derosier DJ, Roberts T, et al. A comparison of subjective and mathematical estimations of fetal heart rate variability. *J Matern Fetal Neonatal Med*. 2008;21(2):101-104.

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: 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;00:1-16. doi:10.1111/aogs.14591

16000412, 0, Downloaded from https://obgyn.onlinelibrary.wiley.com/doi/10.1111/aogs.14591 by Royal Danish Library, Wiley Online Library on [29/06/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License