International Scholars Journals

International Journal of Obstetrics and Gynecology ISSN 2326-7234 Vol. 6 (3), pp. 211-218, March, 2018. Available online at www.internationalscholarsjournals.org © International Scholars Journals

Full Length Research Paper

Interobserver agreement and reliability in intrapartum cardiotocography interpretation, using the 2015 FIGO consensus guidelines

Louise PARET[1,2], Virginie EHLINGER[1], Catherine ARNAUD[1] and Christophe VAYSSIERE[1,2]

[1]Inserm, UMR1027, team SPHERE, Toulouse, France.
[2]Département de gynécolgie-obstétrique, hôpital Paule-de-Viguier, CHU de Toulouse, France.

Accepted 05 December, 2017

To assess interobserver reliability and agreement in the classification of cardiotocograms (CTGs) according to the 2015 revision of the FIGO (International Federation of Gynecology and Obstetrics) CTG classification. Six observers (3 obstetricians and 3 midwives) applied the 2015 FIGO guidelines and independently interpreted 60 intrapartum CTGs, randomly selected after stratification for arterial umbilical pH (pHa) at birth: 1/3 with pHa>7.15, 1/3 with pHa=7.05-7.15, and 1/3 with metabolic acidosis defined by pHa<7.05 and base deficit>10 mmol/L. Interobserver reliability was assessed by the Fleiss kappa coefficient and interobserver agreement by the proportion of agreement (Pa), calculated according to Grant. The overall interobserver reliability was good (kappa=0.62, 95%CI=0.52-0.73). Interobserver agreement was good for CTGs classified as normal by the observers (Pa=71.5%, 95%CI=67.5-75.2), moderate for those classified as pathological (Pa=57.4%, 95%CI=51.3-63.5) and poor for those classified as suspicious (Pa=36.4%, 95%CI=30.9-41.9). Interobserver reliability was good for baseline, moderate for variability assessment and presence of decelerations, but poor for classifying the decelerations. Results did not differ significantly between obstetricians and midwives. Application of this FIGO classification produced good interobserver reliability and agreement on CTG classification overall, but poor reliability for suspicious CTGs and for determination of the type of deceleration.

Keywords: Fetal heart rate, cardiotocography, interobserver variability, interobserver agreement, reliability, reproducibility.

ABBREVIATIONS

CI: Confidence intervals
CTG: Cardiotocography
FIGO: International Federation of Gynecology and Obstetrics
GA: Gestational age
NICHD: National Institute of Child Health and Human Development
Pa: Proportion of agreement

INTRODUCTION

The purpose of monitoring continuous intrapartum cardiotocography (CTG) is to prevent hypoxia/acidosis, which can cause short- and long-term complications, such as hypoxic-ischemic encephalopathy or cerebral palsy. Despite its disappointing absence of effects on the incidence of overall perinatal mortality or cerebral palsy(reported in the Cochrane review published by Alfirevic et al. (2013)), use of continuous intrapartum CTG has become the standard of care in many countries.

Interpretation of CTG abnormalities is not easy. It requires — simultaneously and continuously during labor

Corresponding author E-mail: louise.paret@gmail.com; Tel. +33567771282

— a comprehensive evaluation of various parameters, including baseline rate, variability, reactivity, and presence and type of decelerations. Taking the clinical context into account, clinicians must make the right decision at the right time based on their interpretation of CTG abnormalities: to wait, to correct reversible causes, or to expedite birth, by instrumental or cesarean delivery. Several CTG classifications currently provide clinicians with standardized criteria for assessing CTG characteristics.

Lack of interobserver reliability in CTG interpretation may partly explain why CTG so often fails to predict acidosis at birth. Certainly, despite the use of the 1985 FIGO (International Federation of Gynecology and Obstetrics) classification, interobserver reliability has been assessed at poor to fair (Ayres-de-Campos et al., 1999). Similarly, reproducibility of the NICHD (National Institute of Child Health and Human Development) CTG classification is reported to be poor (Blackwell et al., 2011).

FIGO updated their classification in 2015 to simplify it (Ayres-de-Campos et al., 2015). In this new classification, CTG are categorized as normal, suspicious, or pathological. A suspicious CTG is defined as neither normal nor pathological. Decelerations are tolerated in normal CTG traces if they are neither prolonged nor repeated (Table1).

To our knowledge, the reproducibility of the 2015FIGO CTG classification has been evaluated twice, and both studies found fair interobserver reliability with kappa=0.39, 95%CI 0.33-0.45 for Rei et al. (2016)and kappa=0.38 for Bhatia et al. (2017).

The primary objective of our study was to assess both the interobserver agreement and reliability of the categorization of intrapartum CTG with this classification. The secondary objective was to assess the performance of observers using this CTG classification to predict metabolic acidosis.

MATERIALS AND METHODS

This study took place in March 2016 in a tertiary maternity unit (in Toulouse, in southwestern France) with almost 5000 deliveries each year.

Observations and sampling method

CTG traces of singleton pregnancies were selected in a prespecified population: gestational age (GA) more than 37 weeks, a vaginal delivery or a cesarean section during labor, at least 60 minutes of CTG traces before the pushing stage or before a decision to perform a cesarean. Arterial and venous umbilical cord pH measurements were routinely available. Exclusion criteria were: termination of pregnancy for medical reasons, stillbirth, or severe congenital malformation.

In total, 60 CTG traces were randomly extracted from our CTG database, which contains all CTG traces performed in our unit between June 2013 and March 2015 (n=10,146). Because the statistical tools that evaluate reliability require equivalent numbers of traces across the spectrum of potential neonatal acidosis (none to severe), sampling was planned to over represent pathological traces. We therefore stratified for arterial umbilical pH,with1 in 3 traces with pHa>7.15 (85.4% of the births in our database), 1/3 with pHa between 7.05 and 7.15 (14.2% of births), and 1/3 with metabolic acidosis, defined by pHa<7.05 and a base deficit >10 mmol/L(0.4% of births). CTG traces were anonymized and printed (at a paper speed of 2 cm/min, according to our routine practice) without any annotation. Six observers were asked to analyze these 60 traces(at least 15 months after the birth, to avoid any recall bias).

We considered that 6 observers and 60 CTG traces, with a sufficient number of traces per severity level, provided an appropriate sample size for this study, similar to or higher than most previous studies assessing the reliability of CTG interpretation. To our knowledge, there is no consensual method for calculating power in reproducibility studies that use kappa statistics (Kottner et al., 2011). In addition, 30seemed to be the maximum number of traces that a given observer could assess during a session. Of the six observers, three were midwives and three obstetricians. All had more than two years of experience working in a tertiary center.

Rating Process

The observers first participated in a 1-hour training session (by LP) in using the 2015FIGO CTG classification, with the official FIGO PowerPoint presentation including definitions and case discussions. At the time of the study, the French CTG classification (Carbonne et al., 2013) was routinely used in our center.

Each observer analyzed CTG traces independently during two distinct sessions (30 CTG traces per session, separated by 1 to 20 days) in LP's presence. The order of the traces was randomly selected and was the same for all observers. The 2015 FIGO guidelines were provided to all observers in French (translation by the principal investigator, back-translated into English by a native English-speaking obstetrician for validation). Clinical information was not initially provided. For each trace, the observer had to analyze the last 30 minutes before the pushing stage or before the decision to perform a cesarean. Because the interpretation of variability in CTG requires a minimal length of record, the last 60 minutes of each trace were printed. First, the observers documented each component according to the classification(baseline, variability, reactivity, and presence and type of decelerations) and then classified each CTG tracing as normal, suspicious, or pathological.

Next, clinical details about labor were provided, and the trace during the pushing stage was shown for vaginal

Table 1. 2015 FIGO CTG classification (Ayres-de-Campos et al., 2015).

	Normal	Suspicious	Pathological
Baseline	110-160 bpm [a]	Lacking at least one characteristic of normality, but with no pathological features	<100 bpm[a]
Variability	5-25 bpm [a]		Reduced variability
			Increased variability
			Sinusoidal pattern
Decelerations	No repetitive decelerations		Repetitive late or prolonged decelerations for > 30 min (or > 20 min if reduced variability)
			Deceleration > 5 min
Interpretation	No hypoxia/acidosis	Low probability of hypoxia/acidosis	High probability of hypoxia/acidosis
Clinical management	No intervention necessary to improve fetal oxygenation state	Action to correct reversible causes if identified, close monitoring, or adjunctive methods	Immediate action to correct reversible causes, adjunctive methods or if this is not possible expedite delivery.
			In acute situations, immediate delivery should be accomplished

(a) bpm: beats per minute.

deliveries. The observers were asked to predict whether or not the neonate was born with metabolic acidosis.

Statistical tools

All analyses were performed with Stata version 14 (Stata-Corp. Stata Statistical Software. Release 14, College Station, TX, USA).

To study interobserver reliability, we calculated Fleiss kappa statistics with linear weighting (command kappa2 in Stata (Lazaro et al., 2015)). The kappa coefficient evaluates observed agreement beyond that expected by chance. The 95% confidence intervals (CI) were estimated with the jackknife method(Roberts and McNamee, 2005). Interobserver reliability was categorized as follows: 0–0.20, poor agreement; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, good; and 0.81–1.00, very good. Comparisons of kappa values are reported with their 95% CI.

The proportion of agreement (Pa) between observers was estimated with a method proposed by Grant (1991). Briefly, the Pa between two observers for a category X is the ratio between the number of CTG traces classified “X” by both observers and the number classified “X” by at least one observer. Because each trace is assessed by 6 observers, we considered 60×15 pairs of observers, for 900 trials. The Chi-2 test (or Fisher’s exact test where appropriate) was used for Pa comparisons.

Statistical analysis

First, overall reliability and agreement were calculated. Then, the reliability and agreement of each session were calculated and compared (to explore a possible learning effect). Reliability and agreement were separately calculated for obstetricians and midwives and compared. In addition, we calculated the reliability of the evaluation of each component of the CTG. Finally, sensitivity and specificity for the prediction of metabolic acidosis were calculated for each observer and then compared between observers with the Mc Nemar test for paired data.

RESULTS

Figure 1 reports the distribution of the interpretation of CTG by the 6 examiners. CTG traces were classified as normal in 45% to 52% of cases, suspicious in 13% to 30%, and pathological in 17% to 32% of cases;0% to 13% were judged nonclassifiable.

The overall interobserver reliability was good (kappa=0.62, 95%CI0.52-0.73) (Table 2). Agreement seemed good for CTGs classified as normal by the observers (Pa=71.5%, 95%CI67.5-75.2), moderate for those rated pathological (Pa=57.4%, 95%CI51.3-63.5), and poor for those rated suspicious (Pa=36.4%, 95%CI30.9-41.9) (Table 2).

Interobserver reliability was not significantly different between the first 30 CTGs (first session) and the next 30 (second session). The proportion of agreement was significantly better in the second session than in the first session only for the CTG traces classified as pathological (Table 2).

Fig. Distribution of CTGs classified into FIGO 2015 categories by the 6 observers.

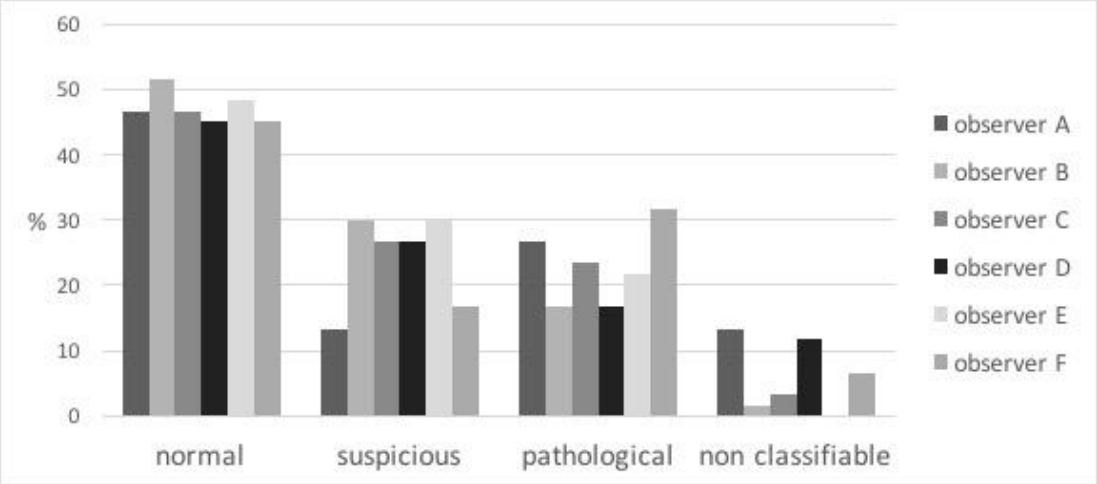

Table 2. Proportion of agreement and kappa coefficients, overall, by session of interpretation, for obstetricians and midwives.

		Normal	Suspicious	Pathological	Kappa
Overall	Trials [a]	484	294	251	
	Pa (%)	71.5	36.4	57.4	K=0.62
	95% CI	(67.5-75.2)	(30.9-41.9)	(51.3-63.5)	(0.52-0.73)
Session 1	Trials [a]	256	172	116	
	Pa (%)	69.1	32.6	47.4	K=0.45
	95% CI	(63.4-74.8)	(25.6-39.6)	(38.3-56.5)	(0.31-0.61)
Session 2	Trials [a]	228	122	135	
	Pa (%)	74.1	41.8	65.9	K=0.59
	95% CI	(68.4-79.8)	(33.0-50.6)	(57.9-73.9)	(0.49-0.72)
P value [b]		0.224	0.106	0.003	
Obstetricians	Trials [a]	96	58	49	
	Pa (%)	76	34.5	57.1	K=0.44
	95% CI	(67.5-84.5)	(22.3-46.7)	(43.2-71.0)	(0.32-0.57)
Midwives	Trials [a]	98	60	53	
	Pa (%)	65.3	35.0	52.8	K=0.63
	95% CI	(55.9-74.7)	(22.9-47.1)	(39.4-66.2)	(0.53-0.74)
P value [c]		0.101	0.953	0.662	

(a) for each CTG (60), number of pairs of observers (maximum 15) with at least one observer placing the CTG in the category under consideration.
(b) comparison of Pa between the 2 sessions by Chi-2 test
(c) comparison of Pa between obstetricians and midwives by Chi-2 test

Interobserver reliability and agreement did not differ significantly between obstetricians and midwives (Table 2).

As Table 3 shows, interobserver reliability was good for the baseline assessment but moderate for that of variability. Reliability was moderate for the presence of decelerations, but ranged from poor to moderate depending on the type of deceleration.

Sensitivity and specificity for predicting metabolic acidosis (20/60 deliveries) with the 2015 FIGO CTG classification varied between observers, ranging respectively from 25.0% (95%CI6.0-44.0) to 47.1% (95%CI23.3-70.8) and from 85.0% (95%CI73.9-96.1) to 97.5% (95%CI92.7-100.0)(Table 4).

DISCUSSION

Overall interobserver reliability for interpretation of 60 CTGs by 6 observers using the new FIGO classification was good. Interobserver agreement was good for CTG traces judged normal, moderate for those assessed as pathological, and poor for those considered suspicious.

Table 3. Interobserver reliability of CTG classification (overall and for each component) according to the 2015 FIGO classification, kappa coefficient and 95% CI.

	kappa	95%CI kappa
Overall	**0.62**	**(0.52-0.73)**
Baseline	0.63	(0.49-0.79)
Variability	0.49	(0.32-0.66)
reduced variability	0.60	(0.41-0.81)
normal variability	0.61	(0.43-0.81)
Presence of decelerations	0.52	(0.38-0.69)
early decelerations	0.19	(0.08-0.33)
variable decelerations	0.38	(0.25-0.52)
late decelerations	0.53	(0.37-0.70)
prolonged decelerations	0.46	(0.31-0.63)
Repetitive decelerations	0.46	(0.30-0.63)

Table 2. Sensitivity and specificity of prediction of metabolic acidosis[a] by observer.

	Sensitivity (%)	95%CI	Specificity (%)	95%CI
Observer A	45.0	(23.2-66.8)	97.3	(92.2-100.0)
Observer B	30.0	(9.92-50.1)	97.4	(92.5-100.0)
Observer C	40.0	(18.5-61.5)	87.5	(77.2-97.7)
Observer D	47.1	(23.3-70.8)	94.6	(87.3-100.0)
Observer E	25.0	(6.0-44.0)	97.5	(92.7-100.0)
Observer F	45.0	(23.2-66.8)	85.0	(73.9-96.1)

(a) defined as arterial umbilical pH<7.05 and base deficit>10mmol/l.

This study has some limitations. First, the participants used the French classification (Carbonne et al., 2013) in their daily practice. Moreover, the training in the FIGO guidelines took place during a single 1-hour meeting. This lack of experience using the FIGO classification might have affected the results and resulted in potential underestimation. However, interobserver reliability did not differ significantly between the two sessions, that is, no learning effect was observed. Second, despite analysis of a total of 360 interpretations (60 CTGs, 6 observers), confidence intervals were wide; they may reflect a lack of power and reduce the impact of our results.

This study has also some strengths. The interpretation sessions were conducted in a standardized manner for each observer, and traces were randomly selected in a database with selection criteria. Another strength was the standardization of the training.

Like many other studies aiming to estimate the interobserver reproducibility of CTG interpretation (Blackwell et al., 2011; Vayssière et al., 2009; Westerhuis et al., 2009), we chose to stratify the CTG traces to be analyzed according to neonatal umbilical pHa at delivery, to increase the proportion of suspicious and pathological CTG. Without stratification before random selection, most of the traces would have been classified as normal, and the Kappa coefficient would have been quite a bit lower, due to the high level of agreement expected by chance.

Previous studies of CTG agreement have varied widely in the criteria chosen to assess reliability. Like some other authors (Bhatia et al., 2017; Blackwell et al., 2011; Rei et al., 2016; Santo et al., 2017), we chose as our primary outcome the overall interpretation of the CTG. Other studies have focused more on the decision about intervention during labor (Amer-Wåhlin et al., 2005;Ojala

et al., 2008; Palomäki et al., 2006; Ross et al., 2004; Vayssière et al., 2009).

Because this outcome depends on a clinical decision based on the CTG interpretation, it appears more useful from a clinical perspective than the CTG classification alone (with its different components). It also, however, appears to be a composite criterion that can result in substantial heterogeneity, due to the variety of interventions possible: correction of reversible causes, close monitoring, adjunctive procedures, or expedited delivery. Moreover, comparison between two decisions to intervene at different times might be difficult: in acute situations, delaying the decision to intervene can increase the risk of severe acidosis and ischemic encephalopathy.

Although the kappa coefficient and the Pa are the best tools for assessing the reliability of an agreement about the classifications of CTG traces, these statistical tools are not intended to compare different classification systems. We summarize here the results of comparisons of different CTG classification systems, but these comparisons cannot determine which system is the best.

Studies using the 1985 FIGO CTG classification showed that its reliability was poor. In a study of 17 intrapartum traceswith 3 observers, Ayres-de-Campos et al. (1999)reported it had an unweighted kappa of 0.31 (95%CI0.11-0.51) for intrapartum CTG. More recently, Santo et al. (2017), in a study of 151 traces with 21 observers, found its reliability was fair, with a kappa coefficient (Light's kappa for n raters) of 0.37 (95%CI0.31-0.43) and moderate agreement (Paranged from 54 to 76%).

Results of the 3-tier NICHD classification did not show it to be clearly better. With120 traces (randomized with stratification for arterial blood pH) interpreted by three observers, Blackwell et al. (2011) found an unweighted Cohen's kappa of 0.45 (moderate reliability, CInot reported). Santo et al. (2017), however, found poor reliability with kappa=0.15 (95%CI0.10-0.21), with poor agreement for categories I (normal, Pa 26%,95%CI18-33) and III (pathological, Pa 26%, 95%CI18-34) but good agreement for category II (suspicious, Pa 83%, 95%CI81-86).

Results for the NICE 2007 classification were equally mediocre with fair reliability (kappa=0.33, 95%CI0.28-0.39) and moderate agreement in all categories (Santo et al., 2017).

Rei et al. (2016) recently published the first study on interobserver reliability and agreement according to the 2015 FIGO classification. Using 151 CTG traces interpreted by 6 observers, they found that agreement was good for traces rated as normal (Pa=67%, 95%CI61-72) and moderate for those rated suspicious and pathological (Pa=54%, 95%CI48-60%, and 59%, 95%CI51-66%, respectively).The overall interobserver reliability was fair (kappa=0.39, 95%CI0.33-0.45), poorer than our results. Results for the Pa were similar to our results except for CTG judged suspicious, which was better than in our study (36.4%, 95%CI 30.9-41.9).

In another study, Bhatia et al. (2017)had 21 observers interpret 10 CTG traces. They found poor reliability (kappa=0.38),but the insufficient description of the type of kappa coefficient used and the absence of CI calculations limits its interpretation (Kottner et al., 2011).

Reliability in our study was good for baseline and moderate for variability and presence of deceleration. However, despite its importance for the global interpretation of CTG,we found that reliability for the type of deceleration was poor. These results are consistent with those of Rei et al. (2016), who found that reliability was moderate for baseline but poor for variability, presence of decelerations, and determination of both variable and late decelerations.Difficulties in interpreting decelerations in our study may explain the lack of reliability of the global interpretation.

We found no significant systematic differences in reliability or agreement between the obstetricians and the midwives, although a trend suggested higher reliability among the midwives. Schiermeier et al. (2011) compared agreement of 24 obstetricians and 19 midwives for 12 CTG traces. They reported a trend toward better agreement among the obstetricians, but no significant results. In any case, as in our study, the observers were volunteers, they were not randomized as a representative sample, and these results cannot be generalized to all obstetricians and midwives.

CTG interpretation is aimed much more at preventing the consequences of asphyxia than predicting normal neonatal outcome. The sensitivity in our study for the prediction of metabolic acidosis was poor, consistently with the results by Chauhan et al. (2008) who found a sensitivity of 0% for predicting umbilical artery pH<7.00. Our observers, however, did not know the proportion of metabolic acidosis in the sample, which was predefined at 33%and thus very different from the proportion they deal with in real life (0.4% in our database).Had they known that the percentage of cases with metabolic acidosis was so unusual, their detection of CTG trace anomalies might well have been much better.

In conclusion, despite good overall inter-observer reliabilityof CTG traces interpreted according to the 2015 FIGO classification, agreement for suspicious CTGs remained poor. A major goal of the 2015 FIGO classification was to try to simplify interpretation, but our study shows that CTG analysis remains a difficult challenge for all practitioners working in delivery rooms. The inadequate performance of CTG for predicting acidosis at birth might be due in part to our focus on pattern recognition rather than a more physiological approach to CTG abnormalities. Teaching a more physiological analysis of intrapartum CTG traces may soon play a role in improving prediction of metabolic acidosis (Chandraharan, 2017).

AUTHOR CONTRIBUTIONS

CV and CA provided LP with advice and counseling for the study design. LP planned and conducted it.
VE provided LP with advice and counseling for the data analysis.
CV and CA helped LP to write the manuscript.

ACKNOWLEDGMENTS

Thanks to the practitioners: Drs Mickaël Allouche, Marion Groussolles, Béatrice Guyard-Boileau and Clémentine Christoph, Jessica Frede, Kévin Ringot.
Thank you to Dr Chloé Dimeglio for her advice, Dr Tracy Chapman for the back-translation and Jo Ann Cahn for English correction.

CONFLICT OF INTEREST

None.

REFERENCES

Alfirevic Z, Devane D, Gyte GML (2013). Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst. Rev. 5:CD006066.

Amer-Wåhlin I, Ingemarsson I, Marsal K, Herbst A (2005). Fetal heart rate patterns and ECG ST segment changes preceding metabolic acidaemia at birth. BJOG Int. J. Obstet. Gynaecol. 112:160–165.

Ayres-de-Campos D, Bernardes J, Costa-Pereira A, Pereira-Leite L (1999). Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision. Br. J. Obstet. Gynaecol. 106:1307–1310.

Ayres-de-Campos D, Spong CY, Chandraharan E, FIGO Intrapartum Fetal Monitoring Expert Consensus Panel (2015). FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography. Int. J. Gynaecol. Obstet. Off. Organ. Int. Fed. Gynaecol. Obstet. 131:13–24.

Bhatia M, Mahtani KR, Nunan D, and Reddy A (2017). A cross-sectional comparison of three guidelines for intrapartum cardiotocography. Int. J. Gynecol. Obstet.

Blackwell SC, Grobman WA, Antoniewicz L, Hutchinson M, Gyamfi Bannerman C (2011). Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System. Am. J. Obstet. Gynecol. 205:378.e1-5.

Carbonne B, Dreyfus M, Schaal J-P, Groupe d'experts des RPC sur la surveillance fœtale au cours du travail (2013). [CNGOF classification of fetal heart rate: color code for obstetricians and midwives]. J. Gynécologie Obstétrique Biol. Reprod. 42:509–510.

Chandraharan E (2017). Handbook of CTG Interpretation: From Patterns to Physiology. Cambridge University Press.

Chauhan SP, Klauser CK, Woodring TC, Sanderson M, Magann EF, Morrison JC (2008). Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability. Am. J. Obstet. Gynecol. 199:623.e1-5.

Collège National des Gynécologues et Obstétriciens Français (2008). [Methods of fetal surveillance during labor. Guidelines]. J. Gynécologie Obstétrique Biol. Reprod. 37 Suppl 1:S101-107.

Grant J (1991). The fetal heart rate trace is normal, isn't it?: Observer agreement of categorical assessments. The Lancet 337:215–218.

Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M, Streiner DL (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J. Clin. Epidemiol. 64:96–106.

Lazaro J, Zamora J, Abraira V, Zlotnik A, Lazaro J, Zamora J, Abraira V, Zlotnik A (2015). KAPPA2: Stata module to produce Generalizations of weighted kappa for incomplete designs.

Ojala K, Mäkikallio K, Haapsamo M, Ijäs H, Tekay A (2008). Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies. Acta Obstet. Gynecol. Scand. 87:536–540.

Palomäki O, Luukkaala T, Luoto R, Tuimala R (2006). Intrapartum cardiotocography -- the dilemma of interpretational variation. J. Perinat. Med. 34:298–302.

Rei M, Tavares S, Pinto P, Machado AP, Monteiro S, Costa A, Costa-Santos C, Bernardes J, Ayres-De-Campos D (2016). Interobserver agreement in CTG interpretation using the 2015 FIGO guidelines for intrapartum fetal monitoring. Eur. J. Obstet. Gynecol. Reprod. Biol. 205:27–31.

Roberts C, McNamee R (2005). Assessing the reliability of ordered categorical scales using kappa-type statistics. Stat. Methods Med. Res. 14:493–514.

Ross MG, Devoe LD, Rosen KG (2004). ST-segment analysis of the fetal electrocardiogram improves fetal heart rate tracing interpretation and clinical decision making. J. Matern.-Fetal Neonatal Med. Off. J. Eur. Assoc. Perinat. Med. Fed. Asia Ocean. Perinat. Soc. Int. Soc. Perinat. Obstet. 15:181–185.

Santo S, Ayres-de-Campos D, Costa-Santos C, Schnettler W, Ugwumadu A, Da Graça LM, FM-Compare Collaboration (2017). Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines. Acta Obstet. Gynecol. Scand. 96:166–175.

Schiermeier S, Westhof G, Leven A, Hatzmann H, Reinhard J (2011). Intra- and interobserver variability of intrapartum cardiotocography: a multicenter study comparing the FIGO classification with computer analysis software. Gynecol. Obstet. Invest. 72:169–173.

Vayssière C, Tsatsaris V, Pirrello O, Cristini C, Arnaud

C, Goffinet F (2009). Inter-observer agreement in clinical decision-making for abnormal cardiotocogram (CTG) during labour: a comparison between CTG and CTG plus STAN. BJOG Int. J. Obstet. Gynaecol. 116:1081-1087; discussion 1087-1088.

Westerhuis MEMH, van Horen E, Kwee A, van der Tweel I, Visser GHA, Moons KGM (2009). Inter- and intra-observer agreement of intrapartum ST analysis of the fetal electrocardiogram in women monitored by STAN. BJOG Int. J. Obstet.Gynaecol. 116:545–551.