AINO JOURNAL, Vol. 15, 2016, pp. 57-64

Original Research

Interobserver variance in intrapartum cardiotocograph interpretation among 7 perinatal experts

Hitomi KIKUCHI — *Lecturer, Department of Medical Engineering, Aino University, Osaka, Japan*

Tomoaki IKEDA — *Professor, Department of Obstetrics and Gynecology, Mie University Faculty of Medicine, Mie, Japan*

Hiroyuki HORIO — *Professor, Graduate School of Applied Informatics, University of Hyogo, Kobe, Japan*

Abstract

The purpose of this study was to evaluate interobserver differences in intrapartum cardiotocograph (CTG) interpretation among 7 perinatal experts in Japan, particularly with regard to the fetal heart rate (FHR) pattern. Printed cardiotocograms (188 pages) of each 9-minute CTG chart segment of approximately the last 3 hours before delivery of 14 high-risk term pregnant women were presented to each clinician. The Japan Society of Obstetrics and Gynecology (JSOG) Standard 82 FHR pattern classifications and the 5-level risk categories were also printed on each page. Seven of the leading Japanese perinatologists were asked to interpret these cardiotocograms. The participants were requested to mark the appropriate FHR pattern type and risk category. The interobserver agreementwas analyzed using proportions of agreement and kappa scores. The average proportion of perinatologists agreeing with the majority opinion for the 5-level risk category was 74.4±4.5%. The kappa score was 0.667±0.053, which indicates a substantial agreement. The high kappa score revealed a consensus in the interpretation of the 5-level risk category by leading Japanese perinatologists. However, differences in opinion arose in the deceleration category of FHR pattern classification.

Key words: Cardiotocogram, fetal heart rate, FHR pattern classification, interobserver agreement, kappa score

Introduction

Intrapartum cardiotocograph (CTG) monitoring is used to prevent non-reassuring fetal status in most perinatal medical facilities in Japan. However, the effectiveness of CTG monitoring has not been explicitly validated according to the Cochrane reviews. Practical CTG interpretation has two factors in which one is based on the classification of the fetal heart rate (FHR) pattern and the other is subsequently recommanded fetal treatment linked to the 5-level risk category from the FHR pattern classification under the standard guideline. The evaluation of whether CTG monitoring improves outcome of delivery has to address the 2aforementioned factors. Therefore, efforts have been made to standardize the interpretation of the FHR pattern. In Japan, the Perinatology Committee of the Japan Society of Obstetrics and Gynecology (JSOG) has defined the terminology regarding FHR pattern classification.[10] In 2007, Parer & Ikeda presented a framework for the standardized management of intrapartum FHR patterns. In 2008, the Perinatology Committee of JSOG defined the intrapartum management guidelines based on the FHR pattern classification, and in 2010, proposed the final version. This guideline proposed a standardized management protocol with 5-level risk categories to address each FHR pattern, along with the general clinical actions for each risk category. Although the FHR classification and treatment of the fetus have thus been standardized, high inter- and intraobserver variance in CTG reading still occurs during the classification of the FHR pattern, resulting in

decreased reliability of CTG monitoring. Therefore, we studied interobserver differences in CTG reading under the JSOG guidelines.

Materials and Methods

FHR pattern classifications for CTG interpretation based on the JSOG guideline are composed of 3 categories, with baseline, variability, and decelerations, which are further classified into 82 pattern types that exclude clinically insignificant and unnecessary cases (Table 1). In addition, the 5-level risk category was generated to clinically treat fetal risk based on 82 FHR pattern types : level 1 indicates the lowest risk, followed by level 2, 3, 4, and 5, which indicates the highest risk. In a previous study, unsatisfactory interobserver agreement was reported for FHR patterns in the high-risk category. In this study, we measured the degree of interobserver agreement for high-risk category FHR patterns obtained from CTGs of 14 term pregnant women, from a single facility, whose umbilical artery pH at birth was below 7.15. A total of 188 printed pages of each 9-minute CTG chart segment of approximately the last 3 hours before delivery was used. Seven of the leading Japanese perinatologists interpreted the printed CTG patterns. The JSOG Standard 82 FHR pattern type classifications (Table 1) and 5-level risk category (Table 2) were also printed on each page. The participants were instructed to mark the appropriate FHR pattern type and risk category. An "unclassified" FHR pattern type was added to the FHR classification table for uncategorized FHR patterns. The participants were not given any clinical information about the patient, and there was no opportunity for the participants to share information among themselves. After we collected the answers from the participants, we calculated the proportions of agreement and the kappa scores for the variation in the FHR interpretation among the participants. The kappa score measures the degree of agreement in classification, but excludes the agreement that occurs by chance, and is scored as a number between 0 and 1. A higher rate of agreement is represented by a value closer to 1. Table 3 (a) shows the interpretation of the kappa scores based on the criteria of Landis & Koch. Kappa score >0.6 was considered "substantial", i. e., a sufficiently high level of agreement among the observers. Table 3 (b) shows the weighted kappa score by Feinsein, which is used for multi-stage and ordinal scale data. To calculate the weighted kappa score, we selected disagreement coefficients according to the degree of difference in the 5-level risk categorization : 0.9 for a 1-level difference, 0.8 for 2 levels, and 0.4 for 3 levels. The analysis of the interobserver differences in the interpretation of the FHR patterns was based on the following points.

Table 2 5-level risk category of FHR

Level	Designation
1	Normal pattern
2	Benign variant pattern
3	Mild variant pattern
4	Moderate variant pattern
5	Severe variant pattern

1. Analysis of the 5-level risk category

1-1) Proportion of perinatologists agreeing with the majority opinion

The majority opinion for the 5-level risk category was defined as no less than 4 participants having the same opinion for 1printed CTG page. Of the 188 samples, 151 (80.3%) received a majority opinion. We analyzed the proportion of

Table 1 FHR pattern classifications with 5-level risk category

Decelerations		None	Early	Variable		Late		Prolonged	
				Mild	Severe	Mild	Severe	Mild	Severe
Normal baseline variability	Normocardia	1	2	2	3	3	3	3	4
	Tachycardia	2	2	3	3	3	4	3	4
	Mild bradycardia	3	3	3	4	4	4	4	4
	Severe bradycardia	4	4		4	4	4		
Decreased baseline variability	Normocardia	2	3	3	4	3	4	4	5
	Tachycardia	3	3	4	4	4	5	5	5
	Mild bradycardia	4	4	4	5	5	5	5	5
	Severe bradycardia	5	5		5	5	5		
Undetectable baseline variability		4	5	5	5	5	5	5	5
Marked baseline variability		2	2	3	3	3	4	3	4
Sinusoidal pattern		4	4	4	4	5	5	5	5

1 : Level 1, 2 : Level 2, 3 : Level 3, 4 : Level 4, 5 : Level 5

Table 3 Interpretation of kappa scores

(a) Kappa score (Landis & Koch)	
<0.00	Poor agreement
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.80–1.00	Almost perfect agreement

(b) Weighted kappa score (Feinstein)	
<0.40	Poor agreement
0.41–0.74	Fair to good agreement
0.75–1.00	Excellent agreement

perinatologists agreeing with the majority opinion based on a total of 1,057 (151×7) data. Because the weighted kappa analysis cannot be applied to the "unclassified" category data, 756 data were submitted to the weighted kappa analysis.

1–2) Proportions of agreement in each risk category

Agreement with the majority opinion shows how much each participant agrees with the classification of 5-level risk category decided by the majority. In contrast, agreement with the minority opinion is calculated by defining the minority opnion as the agreement of no less than 2 participants. The minority opinion essentially reflects all opinions that are not held by only 1 participant.

1–3) Variation of opinion with risk classification according to the majority opinion

We analyzed the variation of opinion for the 756 data described above.

2. Analysis of FHR pattern classification

2–1) Proportions of agreeing perinatologists and kappa scores for the FHR pattern categories

We analyzed the proportions of agreeing perinatologists and the kappa scores for the variability, baseline, and deceleration categories of the FHR pattern.

2–2) Variation of opinion in the FHR pattern categories according to the majority opinion

We analyzed the differences of opinion in the baseline, variability, and deceleration categories of the FHR pattern according to the majority opinion of the 5-level risk category ; exact matches with the majority opinion were omitted (n=114). The baseline category consists of "normocardia," "tachycardia," "mild bradycardia," and "severe bradycardia." The variability category consists of "normal," "decreased," "undetectable," "marked," and "sinusoidal." The deceleration category consists of "none," "early," "variable," "late," and "prolonged" patterns.

3. Variation of opinion in the FHR pattern categories for data omitted from the majority opinion

We analyzed the differences of opinion in the baseline, variability, and deceleration categories of the FHR pattern for the data omitted from the majority opinion in the 5-level risk category (n=37).

Results

1. Analysis of the 5-level risk category

1–1) Proportion of perinatologists agreeing with the majority opinion

The proportions of perinatologists exactly agreeing with the majority opinion are shown in Table 4. The average proportion of agreement was 74.4±4.5%, and the kappa score was 0.667±0.053, which indicated substantial agreement. In addition, the weighted kappa score with the 756 data described above was 0.692±0.026.

1–2) Proportions of agreement in each risk category

Figure 1 shows the frequency with which the perinatologists agreed with the majority and minority opinions in every risk category. The average rate of agreement with the majority opinion was 76.2±21.1% for level 1, 72.7±10.4% for level 2, 73.7±9.2% for level 3, 64.3±16.2% for level , and 57.1%±49.5% for level 5. The average

Table 4 Proportions of agreement, kappa scores, and weighted kappa scores for exact matches with the majority opinion

Reader	Proportion of agreement (%) (n=1057)	Kappa score (n=1057)	Weighted kappa score (n=756)
A	78.1	0.712	0.745
B	78.1	0.716	0.683
C	73.5	0.655	0.705
D	78.8	0.718	0.685
E	64.9	0.557	0.678
F	72.8	0.652	0.682
G	74.2	0.662	0.665
Average±SD	74.4±4.5	0.667±0.053	0.692±0.026

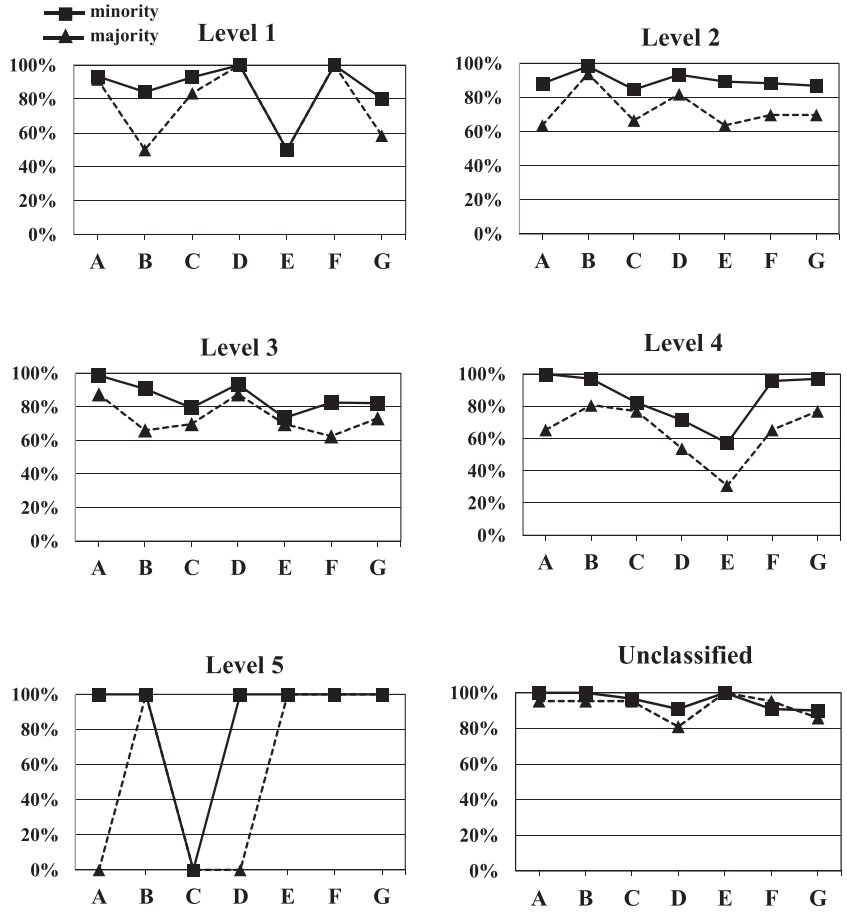

Figure 1 Proportion of exact agreement with the majority opinion (n=151) and the minority opinion (n=188) for each of the 5-level and unclassified risk categories

rates of agreement with the minority opinion ranged from 85.7±35.0% to 95.5±4.4%.

1-3) Variation of opinion with risk classification according to the majority opinion

Figure 2 shows the frequency distribution for the degree of agreement with the majority opinion; "unclassified" data are excluded (n=108). The horizontal axis shows the extent of the difference between the opinion of each reader and the majority opinion, which is set to zero. A positive or negative number indicates that the fetal risk category of the perinatologist was higher or lower than that of the majority opinion, respectively. The vertical axis shows the frequency of the category of agreement; exact matches between the individual opinion and the majority opinion were the most frequent (73.0%), whereas for 24.1% of the data, a one-step difference was observed. Two- and three-step differences were observed for 2.8% and 0.1% of data, respectively.

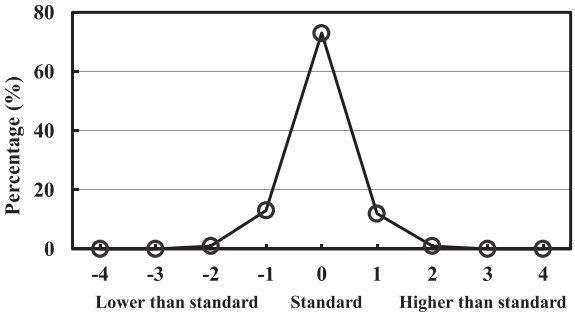

Figure 2 Frequency distribution for the degree of agreement with the majority opinion for the 5-level risk categories

2. Analysis of FHR pattern classification

2-1) Proportions of agreeing perinatologists and kappa scores for the FHR pattern categories

The FHR pattern categories are variability, baseline, and deceleration. The proportions of agreement for the 7 perinatologists for all of the data (n=188) are shown in Figure 3. The agreement was highest in the baseline category (41.0%), with a kappa score of 0.536. The second

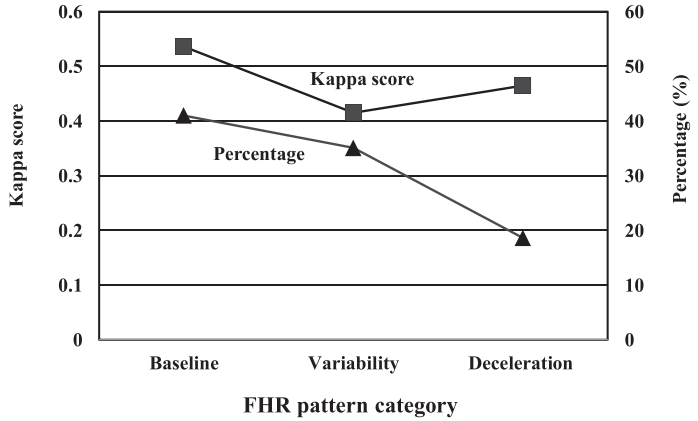

Figure 3 Frequency of exact agreement and kappa scores for each FHR pattern type in the variability, baseline, and deceleration categories

Table 5 Details of variations of opinion with the majority opinions of the 5-level risk category in the variability, baseline, and deceleration categories (n=114)

(a) Baseline category

Baseline category		Number	Percentage (%)
1 selection		80	70.2
2 selections	Normocardia・Tachycardia	29	25.4
	Normocardia・Mild bradycardia	2	1.8
3 selections	Normocardia・Tachycardia・Mild bradycardia	1	0.9
	Normocardi・Mild bradycardia・Severe bradycardia	2	1.8

(b) Variability category

Variability category		Number	Percentage (%)
1 selection		44	38.6
2 selections	Normal・Decreased	28	24.6
	Normal・Marked	25	21.9
	Undetectable・Decreased	3	2.6
	Sinusoidal・Normal	1	0.9
3 selections	Decrease・Normal・Marked	13	11.4

(c) Deceleration category

Deceleration category		Number	Percentage (%)
1 selection		27	23.7
2 selections	Variable・Late	44	38.6
	Late・Prolonged	7	6.1
	None・Variable	7	6.1
	Others	7	6.1
3 selections	None・Variable・Late	7	6.1
	Variable・Late・Prolonged	6	5.3
	Others	5	4.4
4 selections		4	3.5

highest agreement was 35.1% in the variability category, with a kappa score of 0.415. The proportion of agreeing perinatologists was lowest in the deceleration category (18.6%), with a kappa score of 0.465.

2-2) Variation of opinion in the FHR pattern categories according to the majority opinion

To investigate the cause of the disagreement with the majority opnion for the 5-level risk classification (n=114), we analyzed variations in the selection of the 3 categories of the FHR pattern classification— baseline, variability, and deceleration. The results are shown in Table 5 (a) - (c). In the baseline category, Normocardia, Tachycardia, Mild bradycardia, and Severe bradycardia were selected with a frequency of 70.2%, which indicates sufficient agreement with

Table 6 Details of variations of opinion in data that did not reflect the majority opinion in the 5-level risk category among the variability, baseline, and deceleration categories (n=37)

(a) Baseline category

Baseline category		Number	Percentage (%)
1 selection		19	51.4
2 selections	Normocardia・Tachycardia	15	40.5
	Normocardia・Mild bradycardia	1	2.7
	Tachycardia・Mild bradycardia	1	2.7
3 selections	Normocardia・Tachycardia・Mild bradycardia	1	2.7

(b) Variability category

Variability category		Number	Percentage (%)
1 selection		12	32.4
2 selections	Normal・Decreased	15	40.5
	Normal・Marked	4	10.8
	Normal・Sinusoidal	1	2.7
3 selections	Decreased・Normal・Marked	4	10.8
	Decrease・Normal・Sinusoidal	1	2.7

(c) Deceleration category

Deceleration category		Number	Percentage (%)
1 selection		6	16.2
2 selections	Variable・Late	6	16.2
	Late・Prolonged	1	2.7
	None・Variable	1	2.7
	Others	4	10.8
3 selections	None・Variable・Late	10	27.0
	Variable・Late・Prolonged	4	10.8
	Others	2	5.4
4 selections		3	8.1

the majority baseline categorization. However, Normocardia or Tachycardia was selected with a frequency of 25.4%, which reflects the presence of 2 different opinions. Similarly, the frequency of selecting 1 type in the variability category was 38.6% ; furthermore, the frequency of selecting Normal or Decreased was 24.6%, and that of selecting Normal or Marked was 21.9%. The frequency of selecting 1 type in the deceleration category was only 23.7%, and that of selecting Variable or Late was 38.6%.

3. Variation of opinion in the FHR pattern categories for data omitted from the majority opinion

For the 5-level risk category data omitted form the majority opinion (n=37), we also analyzed the variations among the baseline, variability, and deceleration categories of the FHR pattern classification ; the results are presented in Table 6 (a)–(c). Most data in the baseline category were characterized by 1 selection (51.4%) ; however, the frequency of 2 selections of Normocardia or Tachycardia was 40.5%. In the variability category, the frequency of 1 selection was 32.4%, and that of 2 selections, Normal or Decreased, was 40.5%. The opinions in the deceleration category were varied.

Discussion

On the basis of the majority opinion in which at least 4 readers have the same opinion, the average agreement of the perinatologists with the majority opinion in the 5-level risk categoriy was 74.4±4.5%. The kappa score was 0.667±0.053, which indicated substantial agreement. The weighted kappa score was 0.692, which indicates fair to good agreement. As shown in Figure 2, the frequency distributions for the degree of agreement with the majority opinion reveal that when differences arise, the opinion of the perinatologist is usually within 1 step of the majority opinion. Figure 3 reveals that the deceleration category has the lowest agreement ; the proportion of agreement in the deceleration category using 188 samples was 18.6%, and the kappa score was 0.465. The main factor that reduced the rate of

agreement for the 5-level risk category is therefore considered to be the disagreement in the deceleration category of the FHR pattern.

We used kappa scores to determine the rate of agreement. However, if the number of data in each group is biased, the kappa values will be inaccurate. In our study, this bias may have occurred. In Figure 3, the frequency of agreement with the majority opinion in the variability category was 35.1%, and the kappa score was 0.415. In the deceleration category, the frequency of agreement was only 18.6% ; however, the kappa score was 0.465, which was higher than that in the variability category. This discrepancy shows that the kappa score depends on the number of data points. Therefore, it is preferable to evaluate the kappa scores in the context of the frequency of agreement, the intraclass correlation coefficient, and Kendall-s coefficient of concordance. We, therefore, used the kappa scores together with the frequency of agreement in this study.

The proportions of agreement with the minority opinion had high values of 86–95% in each fetal risk category, indicating that at least 2 perinatologists agreed for most of the data. In Figure 1, the agreement of the "unclassified" category of the 5-level risk category and each reader's interpretation ranged from 91.0% to 100%, which indicates that the "unclassified" FHR data were almost always interpreted as "unclassified" by each reader.

In this study of 7 Japanese perinatal experts, the level of agreement with the 2010 JSOG interpretation guideline as determined by a weighted kappa analysis was higher than that in studies investigating the reliability of the FHR interpretation, indicating a consensus in the clinical interpretation of CTG among leading Japanese perinatologists. The same standard CTG interpretation is expected to occur in Japanese medical facilities. However, the deceleration category of the FHR pattern showed differences of opinion ; therefore, there is need for the improvement of the discrimination in this category.

Acknowledgements

We would like to appreciate the contribution given by Dr Takashi Okai, Dr Hiroshi Chisaka, Dr Satoshi Yoneda, Dr Hiroshi Sameshima, and Dr Kunihiro Okamura in interpreting and scoring CTG traces.

Disclosure Statement

None of the authors have any relationships with the companies that may have a financial interest in the information contained in the manuscript nor do they have any conflicts of interest to disclose.

References

American College of Obstetricians and Gynecologists. ACOG Practice Bulletin No. 116 : Management of Intrapartum Fetal Heart Rate Tracings. Obstet Gynecol 116(5) : 1232–1240, 2010

Alfirevic Z, Devane D, Gyte GM : Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst Rev 3 : CD006066, 2013

Ayres-de-Campos D, Spong CY, Chandraharan E, FIGO Intrapartum Fetal Monitoring Expert Consensus Panel : FIGO consensus guidelines on intrapartum fetal monitoring : Cardiotocography. Int J Gynaecol Obstet 131(1) : 13–24, 2015

Blackwell SC, Grobman WA, Antoniewicz L, Hutchinson M, Gyamfi-Bannerman C : Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System. Am J Obstet Gynecol 205 : 378.e1-e5, 2011

Devane D, Lalor J. Midwives : visual interpretation of intrapartum cardiotocographs : intra- and interobserver agreement. J Adv Nurs 52 : 133–141, 2005

Electronic Fetal Heart Rate Monitoring : research guidelines for interpretation. National Institute of Child Health and Human Development Research Planning Workshop. Am J Obstet Gynecol 177 : 1385–1390, 1997

Feinstein AR : Clinimetrics. New Haven, CT : Yale Univ. Press, 1987

Figueras F, Albela S, Bonino S, Palacio M, Barrau E, Hernandez S, Casellas C, Coll O, Cararach V : Visual analysis of antepartum fetal heart rate tracings : inter- and intra-observer agreement and impact of knowledge of neonatal outcome. J Perinat Med 33 : 241–245, 2005

German Society of Gynecology and Obstetrics (DGGG), Maternal Fetal Medicine Study Group (AGMFM), German Society of Prenatal Medicine and Obstetrics (DGPGM), German Society of Perinatal Medicine (DGPM). S1-Guideline on the Use of CTG During Pregnancy and Labor. Geburtshilfe Frauenheilkd 74 (8) : 721–732, 2014

Grivell RM, Alfirevic Z, Gyte GM, Devane D : Antenatal cardiotocography for fetal assessment. Cochrane Database Syst Rev 9 : CD007863, 2015

Hruban L, Spilka J, Chudáček V, Janků P, Huptych M, Burša M, Hudec A, Kacerovský M, Koucký M, Procházka M, Korečko V, Seget-a J, Šimetka O, Měchurová A, Lhotská L : Agreement on intrapartum cardiotocogram recordings between expert obstetricians. J Eval Clin Pract 21(4) : 694–702, 2015

Landis JR, Koch GG : The measurement of observer agreement for categorical data. Biometrics 33 : 159–174, 1977

Liston R, Sawchuck D, Young D : Society of Obstetrics and Gynaecologists of Canada, British Columbia Perinatal Health Program. Fetal health surveillance : antepartum and intrapartum consensus guideline. J Obstet Gynaecol Can 29 (9 Suppl 4) : S3–S56, 2007

Macones GA, Hankins GD, 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. Obstet Gynecol 112 : 661–666,

2008
National Collaborating Centre for Women-s and Children-s Health (UK). Intrapartum care : care of healthy women and their babies during childbirth. NICE Clinical Guidelines 190, 2014
Okai T, Ikeda T, Kawarabayashi T, Kozuma S, Sugawara J, Chisaka H, Yoneda S, Matsuoka R, Nakano H, Okamura K, Saito S, The Perinatology Committee of the Japan Society of Obstetrics and Gynecology : Intrapartum management guidelines based on fetal heart rate pattern classification. J Obstet Gynaecol Res 36 : 925–928, 2010
Parer JT, Ikeda T : A framework for standardized management of intrapartum fetal heart rate patterns. Am J Obstet Gynecol 197 : 26.e1–e6, 2007
Parer JT, Hamilton FE : Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. Am J Obstet Gynecol 203 ; 451.e1–e7, 2010
Rei M, Tavares S, Pinto P, Machado AP, Monteiro S, Costa A, Costa C, Bernardes J, Ayres D : Interobserver agreement in CTG interpretation using the 2015 FIGO guidelines for intrapartum fetal monitoring. Eur J Obstet Gynecol Reprod Biol 205 : 27–31, 2016
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 93 (6) : 596–602, 2014
Sim J, Wright CC : The kappa statistic in reliability studies : use, interpretation and sample size requirements. Phys Ther 85 : 257–268, 2005
The Perinatology Committee report in Japanese. Acta Obstetrica et Gynaecologica Japonica 55 : 1205–1216, 2003 (in Japanese)
The Perinatology Committee report in Japanese. Acta Obstetrica et Gynaecologica Japonica 60 : 1220–1229, 2008 (in Japanese)
Trimbos JB, Keirse MJ : Observer variability in assessment of antepartum cardiotocograms. Br J Obstet Gynaecol 85 : 900–906, 1978