Article

Not peer-reviewed version

Reliability of Cardiotocography in the Assessment of Fetal Heart Rate Patterns

Hitomi Kikuchi * , Tomoaki Ikeda , Nao Murabayashi

Posted Date: 2 June 2025

doi: 10.20944/preprints202506.0033.v1

Keywords: FHR pattern classification; TRIUM; kappa score; inter-rater reliability; intra-rater reliability

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 June 2025 doi:10.20944/preprints202506.0033.v1

Article

Reliability of Cardiotocography in the Assessment of Fetal Heart Rate Patterns

Hitomi Kikuchi [1,*], Tomoaki Ikeda [2] and Nao Murabayashi [3]

[1] Department of Medical Engineering, Aino University, 4-5-4 Higashi-ohda, Ibaraki, Osaka 567-0012, Japan
[2] Department of Obstetrics and Gynecology, Mie University Faculty of Medicine, 2-chōme-174 Edobashi, Tsu, Mie 514-8507 Japan
[3] Department of Reproductive and Perinatal Medicine Hamamatsu University School of Medicine 1-20-1, Handayama, Higashi-ku, Hamamatsu 431-3192, Japan
* Correspondence: h-kikuchi@me-u.aino.ac.jp; Tel.: +81-72-627-1711

Abstract: Background/Objectives: The classification of fetal heart rate (FHR) patterns is performed visually by obstetricians and gynecologists, which sometimes results in problematic discrepancies between assessments. Cardiotocography (CTG), which automatically performs pattern classifications using a computer, has been introduced recently to address this problem. The purpose of this study was to examine the reliability of assessments by obstetricians and by CTG (Trium), as well as identify factors that contribute to differences in assessments. **Methods**: Obstetricians used the FHR guidelines to perform pattern classifications according to baseline, variability, and deceleration of the FHR. Inter-rater reliability, Trium-reliability, and intra-rater reliability were evaluated using Kappa scores. **Results**: There was a high degree of agreement among obstetricians involved in the assessment of FHR patterns. However, deceleration of the FHR was associated with disagreement with the Trium CTG automatic assessment system, and there were clear differences in assessments between mild variable and mild late, as well as severe variable and severe late patterns. **Conclusions**: These findings suggest that the deceleration category program of Trium CTG could be improved to obtain results closer to those of visual assessments.

Keywords: FHR pattern classification; TRIUM; kappa score; inter-rater reliability; intra-rater reliability

1. Introduction

Cardiotocography (CTG) is widely used in clinical practice to continuously measure the fetal heart rate (FHR) and uterine contractions during delivery. In the Japanese guidelines, FHR patterns are classified into 82 types based on three basic patterns (baseline, variability, and deceleration) and their subdivisions. Combinations of FHR patterns are further classified into five levels to estimate the degree of risk for conditions such as fetal hypoxia and acidemia. However, because pattern classification is performed visually by obstetricians and gynecologists, discrepancies are a problem [1–3]. The Trium CTG (Trium Analysis Online GmbH, Munich, Germany), which performs automatic pattern classifications with a computer, has been introduced recently to address this issue. The purpose of this study was to examine the reliability of assessments performed by obstetricians and the Trium CTG, as well as identify factors that contribute to differences in assessments.

2. Materials and Methods

Obstetricians at our hospital followed the FHR guidelines (Tables 1 and 2) [4], recommended by the Perinatal Committee of the Japanese Society of Obstetrics and Gynecology, to perform pattern classifications. The classification of FHR patterns refers to the baseline, variability, and deceleration of the FHR. The data analyzed were obtained from January 2011 to December 2017 and were read by CTG at the National Cerebral and Cardiovascular Center and Mie University Faculty of Medicine which have the same FHR monitors and were read by CTG. Informed consent was provided verbally and the data analyzed retrospectively. The study was approved by the Research Ethics Committees of the Graduate School of Applied Informatics, University of Hyogo and Aino University (approval number: Aino2012–010).

Table 1. The 5-level classification of FHR patterns (82 categories) [4].

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	4	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. FHR, fetal heart rate.

Table 2. Levels of FHR patterns [4].

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

FHR, fetal heart rate.

During the assessments, the clinical background of the patients was not provided, nor was there any opportunity to exchange this kind of information among the raters. We added “assessment not possible” for cases to which the guidelines did not apply. The following analyses were conducted to examine the reliability of the assessments and factors associated with differences: inter-rater reliability, Trium reliability, and intra-rater reliability.

The Kappa score was used to evaluate reliability. Kappa scores were categorized using Landis & Koch's assessment table (Table 3) [5].

Table 3. Kappa score (Landis & Koch) [5].

<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
0.81–1.00	Almost perfect agreement

2.1. Inter-Rater Reliability

Five raters from the same institution and five from different institutions, all of whom were obstetricians in leadership positions in Japan, performed FHR pattern classification on 188 CTG sections from 14 pregnant women. Data for which assessment was not possible were excluded and those with agreement of more than half of the raters (three) were included. The Kappa scores of raters from the same and different institutions, as well as Kappa scores by category, were analyzed.

2.2. Trium Reliability

Eleven obstetricians working at the same institution performed FHR pattern classification on 101 CTG sections from 12 pregnant women. Data for which assessment was not possible were excluded and those with agreement of more than half of the raters were included. The Kappa scores between raters and Trium, as well as Kappa scores by category, were analyzed.

2.3. Intra-Rater Reliability

One obstetrician performed a second assessment of the pattern classifications of the 101 sections from the 12 pregnant women after a month had passed since the first assessment. Agreement between the first and second assessments was analyzed.

3. Results

3.1. Inter-Rater Reliability

3.1.1. Kappa Scores of Raters from the Same and Different Institutions

Table 4 shows the Kappa scores of the five raters from the same institution and the five from different institutions. Both had a fair agreement on variability and level. The reference values in the table indicate the degree of agreement between the majority opinions of raters from the same and different institutions. Agreement between them was “substantial" for variability and level, as well as “almost perfect" for baseline and deceleration.

Table 4. Kappa scores of raters from the same and different institutions.

	Same institution	Different institutions	Reference value
Baseline	0.60	0.57	0.87
Variability	0.45	0.31	0.79
Deceleration	0.48	0.56	0.94
Level	0.39	0.38	0.76

3.1.1. Kappa Score by Category

Table 5 shows the Kappa scores of each FHR category. Kappa scores tended to be higher for normal patterns in all categories and for raters from both the same and different institutions. Moderate agreement was seen with tachycardia, and a severe prolonged pattern showed substantial to almost perfect agreement, demonstrating high levels of agreement overall.

doi:10.20944/preprints202506.0033.v1

Table 5. Kappa score categories.

	Category	Different institutions	Same institution
Variability Different institutions: n=46 Same institution: n=146	Moderate	0.32	0.45
	Minimal	0.30	0.46
	Marked	0.12	0
	Sinusoidal	0.12	0
	Absent	0	0
Baseline Different institutions: n=143 Same institution: n=150	Normocardia	0.57	0.61
	Mild bradycardia	0.06	0.29
	Severe bradycardia	0.37	0
	Tachycardia	0.65	0.63
Deceleration Different institutions: n=111 Same institution: n=135	None	0.64	0.70
	Early	0.00	0.09
	Mild variable	0.56	0.45
	Mild late	0.33	0.31
	Mild prolonged	0.18	0.28
	Severe variable	0.09	0.20
	Severe late	0.58	0.45
	Severe prolonged	0.81	0.72
Level Different institutions: n=126 Same institution: n=135	Level 1	0.48	0.66
	Level 2	0.38	0.36
	Level 3	0.38	0.30
	Level 4	0.36	0.36
	Level 5	0.23	0.44

3.2. Trium Reliability

3.2.1. Kappa Score Between Raters, as Well as Between Trium and Raters

The majority of the 11 raters (six or more) agreed on the pattern classifications of the 101 sections in baseline (98 [97%]), variability (96 [95%]), deceleration (61 [60%]), and level (66 [65%]). Kappa scores between the 11 raters and Trium ranged from fair to moderate agreement for all categories. However, agreement with the reference values ranged from substantial to almost perfect in baseline and variability but ranged from moderate in deceleration and level (Table 6).

Table 6. Kappa score between raters, as well as between Trium and raters.

	Assessments of 11 raters	Trium	Reference value
Baseline (n=98)	0.34	0.49	0.94
Variability (n=96)	0.43	0.45	0.71
Deceleration (n=61)	0.55	0.36	0.54
Level （n=66）	0.45	0.38	0.46

3.2.2. Kappa Score by Category

Table 7 shows the Kappa scores of each category. Kappa scores tended to be relatively high for the normal results in all categories, as well as for both raters from the same institution and Trium. Kappa scores of Trium tended to be higher in variability and baseline, while those of the 11 raters from the same institution tended to be higher in deceleration and level.

Table 7. Kappa scores by category.

	Category	11 raters	Trium
Variability	Moderate	0.32	0.59
	Minimal	0.3	0.68
	Marked	0.29	0
	Sinusoidal	0	0
	Absent	0	0
Baseline	Normocardia	0.43	0.80
	Mild bradycardia	0	0
	Severe bradycardia	0	0
	Tachycardia	0.44	0.80
Deceleration	None	0.77	0.74
	Early	0.54	0
	Mild variable	0.54	0.46
	Mild late	0.33	0
	Mild prolonged	0.25	0
	Severe variable	0.56	0.40
	Severe late	0.33	0
	Severe prolonged	0.53	0.37
Level	Level 1	0.67	0.67
	Level 2	0.40	0.31
	Level 3	0.38	0.28
	Level 4	0.24	0.65
	Level 5	0.13	0

3.2.3. Intra-Rater Reliability

Table 8 shows the agreement when one obstetrician performed a second reading sometime after a month reading. The agreement between the two FHR pattern classifications was very high at ≥90% in baseline and variability, in which the Kappa scores were between substantial and almost perfect agreement.

In contrast, the agreement rate in deceleration and level was approximately 60%, and the Kappa scores also showed moderate agreement.

Table 8. Agreement for the same obstetrician (n=101).

	Baseline	Variability	Deceleration	Level
Kappa coefficient	0.93	0.71	0.54	0.45
Agreement rate (%)	99%	93%	60%	60%

4. Discussion

The reliability of CTG readings has been assessed using various methods. One example is to link the CTG results with actual birth outcomes (umbilical cord artery blood data) for analysis, assessment of CTG data by different readers, and checking the agreement rate.

The Japanese Society of Obstetrics and Gynecology has established five levels to estimate the degree of risk of conditions such as fetal hypoxia and acidemia. FHR pattern classification includes 82 categories, based on the results of baseline, variability, and deceleration. However, the number of categories means that it is difficult to perform assessments; therefore, a five-level classification is generally used. This is not seen as a problem because the medical treatment is the same. However, this also means that discrepancies may exist in level assessments. We analyzed inter-rater reliability, Trium reliability, and intra-rater reliability to examine the reliability of FHR pattern classification and factors contributing to discrepancies.

The results of Kappa scores within the same institution and between institutions showed fair agreement on both variability and level. However, the degree of agreement between readers was relatively good, and there were no differences between institutions. The factor contributing to differences in the interpretation between the automatic CTG assessment system (Trium) and obstetricians was identified as decelerations. The reason could be due to differences between the assessment of mild variable and mild late, as well as between severe variable and severe late, which caused the agreement on levels to decrease.

Intra-rater reliability for a single obstetrician is higher than the inter-rater agreement rate between obstetricians [3]. Previous studies have shown that agreement rates among different raters are highly variable [2], which was also the case in the present study. The reason for this was the assessment of deceleration, with the most common disagreements occurring between mild variable and severe variable, as well as between severe variable and severe late. However, we believed this was due, in part, to assessor factors.

In this study, we examined the factors of disagreement in CTG readings. Differences associated with the subjectivity of readers and interpretation of guidelines were some hypothesized factors, but the results of this study negate these hypotheses. Another factor was the quality of CTG, particularly with the automatic CTG assessment system (Trium), where unreadable data is not an option and all readings are classified into some category. Results are always classified into some category even when patterns consist of many indistinguishable complicated patterns in the 10 minutes. In this study as well, mild variable, mild late, severe variable, and severe late sometimes appeared as temporary changes, and these were likely important factors for disagreement between readings.

Kappa scores were used in this study, which are known to be lowered by data bias [6]. Previous publications have shown that agreement rates decline with worse FHR patterns [7–9], which, therefore, was avoided by utilizing CTG from cases with umbilical cord arterial blood pH <7.15 at birth in the current study. Nevertheless, the high percentage of normal waveforms suggested that some Kappa scores may have been lower than they would have normally been. Moreover, one of the limitations of the study was the discrepancy in the results of the automatic CTG assessment system (Trium) due to deceleration; it is suggested to improve this to enhance the agreement on the level.

5. Conclusions

This study found a high rate of agreement in FHR pattern readings between specialist obstetricians. Meanwhile, the assessment of deceleration by the same obstetrician showed only 60% agreement, highlighting the ongoing difficulty in interpreting decelerations.

Author Contributions: Conceptualization: T.I.; Data curation: T.I.; Formal analysis: H.K.; Investigation: N.M.; Methodology: H.K.; Project administration: T.I.; Resources: H.K.; Software: H.K.; Supervision: T.I.; Validation: N.M.; Visualization: H.K.; Writing – original draft: H.K.; Writing – review & editing: H.K.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Research Ethics Committees of the Graduate School of Applied Informatics, University of Hyogo and Aino University (approval number: Aino2012–010).

Informed Consent Statement: Informed consent was provided verbally and the data analyzed retrospectively.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgments: We would like to thank the doctors of the Mie University Hospital Department of Obstetrics for their assistance in performing the CTG assessments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FHR	fetal heart rate
CTG	Cardiotocography

References

1. Devane, D.; Lalor, J. Midwives' visual interpretation of intrapartum cardiotocographs: Intra- and inter-observer agreement. *J Adv Nurs* **2005**, *52*, 133–141. DOI:10.1111/j.1365-2648.2005.03575.x.
2. Blackwell, S.C.; Grobman, W.A.; 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* **2011**, *205*, 378.e1–378.e5. DOI:10.1016/j.ajog.2011.06.086.
3. Hernandez Engelhart, C.; Gundro Brurberg, K.; Aanstad, K.J.; Pay, A.S.D.; Kaasen, A.; Blix, E.; Vanbelle, S. Reliability and agreement in intrapartum fetal heart rate monitoring interpretation: A systematic review. *Acta Obstet Gynecol Scand* **2023**, *102*, 970–985. DOI:10.1111/aogs.14591.
4. Okai, T.; Ikeda, T.; Kawarabayashi, T.; Kozuma, S.; Sugawara, J.; Chisaka, H.; Yoneda, S.; Matsuoka, R.; Nakano, H.; Okamura, K.; et al. Intrapartum management guidelines based on fetal heart rate pattern classification. *J Obstet Gynaecol Res* **2010**, *36*, 925–928. DOI:10.1111/j.1447-0756.2010.01342.x.
5. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. DOI:10.2307/2529310.

6. Tsushima, E. Application of coefficient of reliability to the studies of physical therapy. *Rigakuryoho Kagaku* **2002**, *17*, 181–187. DOI:10.1589/rika.17.181.
7. Bernardes, J.; Costa-Pereira, A.; Ayres-de-Campos, D.; van Geijn, H.P.; Pereira-Leite, L. Evaluation of interobserver agreement of cardiotocograms. *Int J Gynaecol Obstet* **1997**, *57*, 33–37. DOI:10.1016/s0020-7292(97)02846-4.
8. Ayres-de-Campos, D.; Bernardes, J.; Costa-Pereira, A.; Pereira-Leite, L. Inconsistencies in classification by experts of cardiotocograms and subsequent clinical decision. *Br J Obstet Gynaecol* **1999**, *106*, 1307–1310. DOI:10.1111/j.1471-0528.1999.tb08187.x.
9. Epstein, A.J.; Twogood, S.; Lee, R.H.; Opper, N.; Beavis, A.; Miller, D.A. Interobserver reliability of fetal heart rate pattern interpretation using NICHD definitions. *Am J Perinatol* **2013**, *30*, 463–468. DOI:10.1055/s-0032-1326991.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.