OBSTETRICS

Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation

Julian T. Parer, MD, PhD; Emily F. Hamilton, MD

OBJECTIVE: The purpose of this study was to measure agreement among 5 expert clinicians and a computerized method with the use of a strict fetal heart rate classification method.

STUDY DESIGN: Five providers independently scored 769 8-minute segments from the last 3 hours of 30 tracings with the use of a 5-tier color-coded framework that contains pattern descriptions and proposals for management. Computer analysis was performed with PeriCALM Patterns (PeriGen, Princeton, NJ) to detect and classify patterns.

RESULTS: The clinicians agreed exactly with the majority opinion in 57% (95% confidence interval [CI], 49–64%) of the segments and were within 1 color code in 89% (95% CI, 81–96%). The average proportion of agreement was 0.83 (95% CI, 0.73–0.94). Weighted Kappa scores averaged 0.58 (range, 0.48–0.68). The computer-based results were not statistically different: 0.87 and 0.52, respectively.

CONCLUSION: These 5 clinicians achieved moderate-to-substantial levels of agreement overall using a strictly defined method to classify fetal heart rate tracings. The result of the computerized method was similar to the conclusions of these clinicians.

Key words: computer, electronic fetal heart rate monitoring, interobserver agreement

Cite this article as: Parer JT, Hamilton EF. Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010;203:●●●●.

A major limitation of electronic fetal heart rate monitoring (EFM) interpretation has been an unacceptably high inter- and intraobserver variation in interpretation.[1-7] Such variation hampers the important clinical goals of accurate communication and application of timely management.[8-12] Recent efforts to lessen the problems of delayed intervention for abnormal tracings have resulted in a variety of methods to categorize tracings and guide management.[13-19] The rationale underlying this movement is based, in part, on the premise that these explicit definitions will enable clinicians to categorize tracings more consistently.

We chose to evaluate clinical performance using a 5-level classification method.[19] Multiple levels ensure that each level spans a smaller range of severities, compared with a simpler classification in which very disparate subgroups could be grouped together in a level. Multilevel classification methods are useful clinically but are challenging to apply consistently, especially when there are many factors to consider and the task must be done repeatedly under conditions of fatigue and distraction. The classification method was based on 134 different combinations of fetal heart rate (FHR) characteristics. The characteristics of the FHR patterns were defined rigidly, as were the combinations that comprise each level. In addition, each level of the framework was linked to a different proposal for management.

Previous investigators who used a variety of approaches to measure clinician agreement, such as comparing specific characteristics of the tracings (eg, type of deceleration, quantity of variability, or combinations of these features), have shown very poor levels of agreement.[1-7] Furthermore, none of these studies used such a complex classification schema. Thus, it is very pertinent to determine whether such a classification method actually could help clinicians to achieve consistency in EFM interpretation.

In addition, we sought to determine how well a computerized version of this method would compare to the clinicians. PeriCALM Patterns (PeriGen, Princeton, NJ) is a validated Food and Drug Administration–cleared software package that identifies and measures FHR baseline, baseline variability, and accelerations and decelerations based on the National Institute of Child Health and Human Development definitions.[13] The computerized method with the use of this software and the 5-level classification schema previously was subjected to independent testing for discriminating capacity in a series of 2132 tracings from deliveries that covered a wide range of outcomes. There was a clear correlation between the severity and duration of aberrant FHR patterns and newborn infant state.[20]

Materials and Methods

This multiple reader/multiple case study design included 5 clinical experts, specialized software for FHR analysis, and EFM records from 30 singleton term labors. The cases all had umbilical artery

From the Department of Obstetrics, Gynecology and Reproductive Sciences, University of California, San Francisco, School of Medicine, San Francisco, CA (Dr Parer); the Department of Obstetrics and Gynecology, McGill University, Montreal, Quebec, Canada; and PeriGen, Princeton, NJ (Dr Hamilton).

Received Jan. 15, 2010; revised March 15, 2010; accepted May 19, 2010.

Reprints not available from the authors.

Authorship and contribution to the article is limited to the 2 authors indicated. There was no outside funding or technical assistance with the production of this article.

0002-9378/$36.00

doi: 10.1016/j.ajog.2010.05.037

TABLE 1
Summary of the 5-level classification

	Decelerations		Recurrent variable			Recurrent late			Prolonged		
Variability (baseline)	**None**	**Early**	**Mild**	**Moderate**	**Severe**	**Mild**	**Moderate**	**Severe**	**Mild**	**Moderate**	**Severe**
Moderate (normal)											
Tachycardia	B	B	B	Y	0	Y	Y	0	Y	Y	0
Normal	G	G	G	B	Y	B	Y	Y	Y	Y	0
Mild bradycardia	Y	Y	Y	Y	0	Y	Y	Y	Y		0
Moderate bradycardia	Y	Y			0		0	0			0
Severe bradycardia	0	0			0			0			0
Minimal											
Tachycardia	B	Y	Y	0	0	0	0	R	0	0	R
Normal	B	B	Y	0	0	0	0	R	0	0	R
Mild bradycardia	0	0	R	R	R	R	R	R	R	R	R
Moderate bradycardia	0	0			R		R	R			R
Severe bradycardia	R	R			R			R			R
Absent											
Tachycardia	R	R	R	R	R	R	R	R	R	R	R
Normal	0	R	R	R	R	R	R	R	R	R	R
Mild bradycardia	R	R	R	R	R	R	R	R	R	R	R
Moderate bradycardia	R	R			R		R	R			R
Severe bradycardia	R	R			R			R			R

B, blue(2); *G*, green(1); *O*, orange(4); *R*, red(5); *Y*, yellow(3).

Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

blood gases evaluated at birth and spanned a range of newborn infant outcomes and complexity of FHR patterns. The tracings covered the last 3 hours before birth. They were reproduced in their original size and assembled in booklets with 8 minutes of tracing per page. A total of 769 pages were presented to each clinician. Unknown to the clinicians, 13 of these tracings came from babies with elevated umbilical artery base deficit values at birth (>12 mmol/L) and encephalopathy in the early neonatal period.

Five obstetric providers, 4 perinatologists, and 1 certified nurse midwife, all of whom were clinically active and have been published in FHR monitoring literature, were recruited to score each page in the booklets according to the 5-tier color-coded system. The practitioners were aware of this classification method previously and used it in clinical practice to varying degrees. Each expert was given a detailed set of instructions and a colored worksheet that outlined the 5-tier framework (Table 1).[19,21] Combinations of the various FHR pattern features (namely baseline rate, variability, accelerations, and decelerations) defined the 5 colors. For example, "green (1)" required all features to be within normal limits. Progressively abnormal combinations of baseline rate, reduction of variability, and increased depth and/or duration of decelerations defined the "blue(2)", "yellow(3)", "orange(4)," and "red(5)" categories.

The clinicians were given a written list of the rules for interpretation and quantification (Table 2) and were encouraged to follow the rules, even if they disagreed with them, because the aim of the project was to measure the levels of agreement when this particular rule-based classification method was being used. The identity of the experts was not divulged, so they had no opportunity to interact among themselves regarding the scoring.

The same tracings were subjected to the computerized method. Some of the FHR conditions in the classification by Parer and Ikeda[19] required additional specification for the computer system. For example, in mathematic terms, absent baseline variability (0 beats/min) would be equivalent to a perfect flat line, which does not exist in living biologic entities. Thus, for this exercise, *absent variability* was defined as a measured variability of <2 beats/min amplitude, and *minimal variability* was defined to be between 2 and 5 beats/min.

We used several techniques to evaluate agreement between the various "readers" and the reference group. A reader refers to a specific clinician or the computerized method that is making the assessment. The reference group never included the reader under evaluation.

First, we measured the performance of each of the 5 clinicians by comparing his/her readings to all the assessments by the

TABLE 2

Definitions and quantitation of fetal heart rate characteristics

Decelerations	Quantitation	Definition
Recurrence of decelerations (at least 2 decelerations and at least 50% of the contractions have associated decelerations in a 20-minute window)		
Variable decelerations are considered	Severe	If they are recurrent and last 1-2 minutes and touch 70 beats/min
		If they are recurrent and last >2 minutes and touch 80 beats/min
	Moderate	If they are recurrent and last 30-60 seconds and touch 70 beats/min
		If they are recurrent and last >60 seconds and touch 80 beats/min
	Mild	All else
Late decelerations	Severe	If they are recurrent and are >45 beats/min below baseline
	Moderate	If they are recurrent and are 15-44 beats/min below baseline
	Mild	If they are recurrent and are <15 beats/min below baseline
Prolonged decelerations (last >2 min but <10 min)	Severe	If they are ≤70 beats/min
	Moderate	If they are down to 70-80 beats/min
	Mild	If they are not <80 beats/min
Bradycardia (baseline level for >10 min)	Severe	If they are ≤70 beats/min
	Moderate	If they are down to 70-80 beats/min
	Mild	If they are not <80 beats/min

The readers received the following instructions: each page contains 8 minutes of tracing; code by the highest risk in each page: 1, that is, severe trumps moderate or mild; 2, late decelerations trump variable decelerations; 3, prolonged decelerations trump variable or late decelerations; 4, minimal or undetectable fetal heart rate variability trumps moderate, if present for >50% of the time.

Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

other clinicians. That is, the 769 assessments of 1 clinician (the reader) were compared with 3076 assessments of the other 4 clinicians (the clinical reference: 769 × 4). In practice, most clinicians left a few assessments blank; therefore, the actual numbers of possible comparisons were slightly less (Table 3). The percentage of agreement indicates how often a reader assigned exactly the same color among all the color assignments made by the reference group. The overall performance of the clinical group was the average and 95% confidence interval [CI] of these 5 readers. The performance of the computerized method was calculated in the same way and then compared with the average and 95% CI of the 5 clinicians.

Second, we were interested to know how each reader performed on specific tracing segments in which there was substantial agreement within the reference group as to color assignment. Therefore, for each clinician, we measured the level of agreement with the majority opinion for that segment. This reduced the number of possible comparisons because each tracing segment could have only 1 majority opinion, and some segments had no majority. The overall performance of the clinical group was the average and 95% CI of these 5 readers. The performance of the computerized method was calculated in the same way and compared with the clinical group.

Third, we were interested to know not only exact agreement levels but also levels of close agreement. *Close agreement* was defined when a reader assigned a color that was within plus or minus 1 color of the majority opinion. We summarized the performance of each clinician and computer as described earlier. In addition, we calculated how often each reader deviated by 1, 2, 3, or 4 levels.

Percentages of agreement show only a partial picture. They do not consider agreement that could have occurred by chance or the impact of very discrepant disagreements. Therefore, we applied 2 standard statistical measures to assess

TABLE 3

Percentages of exact agreement with all clinical opinions

Clinician	Comparisons, n	Agreement, %
E	3029	48.6
A	3018	48.0
D	2933	46.8
B	3029	42.7
C	3017	40.3
Average[a]		45.5 ± 3.6
95% confidence interval		42.1–48.4
Computer	3797	44.9

[a] Data are given as mean ± SD.

Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

TABLE 4
Percentages of exact agreement with the majority opinions

Clinician	Majority comparisons, n	Exact match, %
E	528	66.1
A	536	61.0
D	522	59.8
B	559	51.0
C	598	45.5
Average[a]		56.7 ± 8.3
95% confidence interval		49.4–63.9
Computer	549	56.8

[a] Data are given as mean ± SD.

Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

this. Kappa scores indicate how much the parties agreed beyond the level of agreement that could be expected by chance alone. The proportion of agreement reflects the effect of deviating assessments. It measures how often the reference group agreed with the reader's assignment. If the reference group always agreed with the reader's assessment, the proportion of agreement would be 1.0. This measure is equivalent to a positive predictive value used to measure the performance of a diagnostic test against a gold standard. Both Kappa scores and proportion of agreements can range from 0 (very poor) to 1 (perfect). These latter 2 measures were stringent tests because they examined exact agreements.

The CIs were based on the Student *t* test distributions and assumptions that the results of the readers in the reference group were distributed normally. With only 5 members in the reference group, it was not possible to adequately test for normality. We calculated Kappa scores with linear weighting.[22]

The project was exempt from Committee on Human Research, University of California, San Francisco, review because a code no longer exists that links the data that were used to individual patients.

RESULTS

Table 3 shows the percentage of exact agreement for each clinical reader compared with all other clinical assessments. On average, the percentage of exact agreement among the clinicians was 45.3% (95% CI, 42.1–48.4%). The rate of exact agreement for the computerized method was 44.9% or very similar.

Table 4 shows how frequently each reader agreed with the reference made up of majority opinions. On average, the percentage of exact agreement for the clinicians with the majority opinion was 56.7 % (95% CI, 49.4–63.9%). The percentage of exact agreement for computerized method was similar at 56.8%.

Proportions of agreement and Kappa scores for exact matches to the majority opinion are shown in Table 5. Proportions of agreement were high, and the Kappa scores indicated moderate- to-substantial agreement. Results for the computerized method were similar and lay well within the 95% CI of the clinicians.

The Figure shows greater detail on the amount of disagreement between the reader and the majority opinion reference. The horizontal axis displays the number of color-coded levels between the opinion of the reader and the majority opinion. Performance peaked at 0, which represents an exact match. Clinician assessments were within 1 color code in 88.6% (95% CI, 80.8–96.4). The performance of the computerized method was 83.1% and within the 95% CI of the clinicians.

The data in Table 6 examine performance at each specific color level. Percentages of exact agreement were best when the tracings were clearly normal or "green(1)" The 3 intermediate color levels showed less agreement. The least agreement was found with "red(5)." Segments classified as red(5) occurred infrequently, which resulted in a very small sample size for the assessment of performance at this particular color.

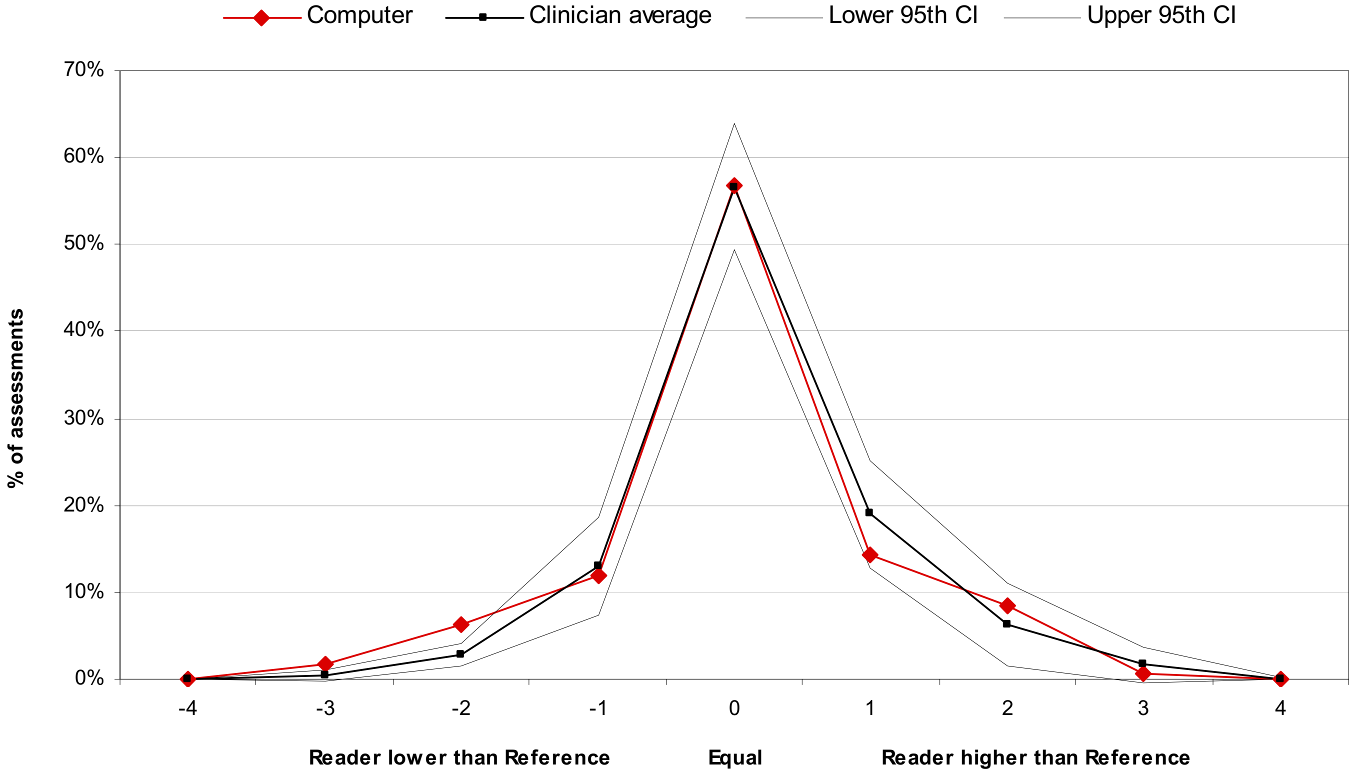
Measurement of close agreement that was defined by agreement within 1 color level was possible only in the middle ranges that were bracketed by another color. Percentages of close agreement to the majority opinions of blue(2), yellow(3), and orange(4) were 85.9%, 85.6%, and 97.3% for the clinicians and 97.9%, 88.2%, and 82.6% for the computerized method, respectively. Only the latter computerized measure fell below the 95% CI for the clinicians.

TABLE 5
Proportions of agreement and weighted Kappa scores for exact agreement with the majority opinion

Clinician	Proportion of agreement	Kappa score
E	0.93	0.67
A	0.88	0.63
D	0.86	0.61
B	0.76	0.53
C	0.73	0.47
Average	0.83	0.58
95% confidence interval	0.73–0.94	0.48–0.68
Computer	0.87	0.52

Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

FIGURE
Frequency distributions for degree of agreement with the majority opinion



Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

COMMENT

We have examined the agreement among 5 select experts who used a strictly defined set of rules. With the use of a variety of methods, all results indicated moderate-to-substantial agreement among the study clinicians for exactly matching the majority opinion of the reference group. Performance measures with the use of a more lenient test, a close match, were even higher.

These levels of agreement are much higher than previously reported.[1-6] For example proportions of agreement previously reported for a simpler task, namely identification of individual features were 0.46 for accelerations and

TABLE 6
Percentages of exact agreement with the majority opinion at each of the 5 color-coded categories

	Majority opinion reference, %				
Clinician	Green(1) [n = 1230]	Blue(2) [n = 413]	Yellow(3) [n = 682]	Orange(4) [n = 358]	Red(5) [n = 60]
E	88.6	43.4	63.2	50.7	0
A	85.7	53.1	32.2	50.0	100.0
D	73.0	48.1	40.8	71.2	0
C	64.2	29.0	23.1	43.4	100.0
B	58.4	46.7	57.6	34.6	7.7
Average[a]	74.0 ± 13.2	44.0 ± 9.1	43.4 ± 16.9	50.0 ± 13.5	41.5 ± 53.5
95% confidence interval	62.5–85.5	36.1–52.0	28.6–58.2	38.1–61.8	–5.3 to 88.4
Computer	75.5	32.3	58.8	40.6	7.4

[a] Data are given as mean ± SD.

Parer. Rule-based fetal heart rate interpretation. Am J Obstet Gynecol 2010.

0.11-0.55 for decelerations.[5,6] The improved performance that we noted may reflect the effects of several recent developments in obstetrics that include a concerted effort by national bodies to promote the use of standard nomenclature with explicit definitions of the FHR terms and increasingly prevalent requirements to demonstrate maintenance of competence in EFM interpretation. Finally, our choice of exceptional readers with recognized expertise in EFM would also tend to produce higher results. That said, the very diverse and large number of tests (>700 per reader) and similar trends in the results no matter what analysis strategy was used indicate the robustness of our conclusions. Contrary to previous reports, we have shown that these clinicians, admittedly experts, can analyze tracings with good levels of agreement.

We have also compared the performance of a computerized method to these select clinicians. The computerized method also performed well, was highly similar to the average performance of the clinical group, and lay well within their 95% CI. The only exception occurred in the small sample of the most abnormal group of tracings, for which the computer identified fewer cases. We speculate that a possible explanation may be that the definition that was used by the software for absent variability was too stringent.

This study was designed to measure degree of agreement, not accuracy. When disagreement occurred, it was not necessarily obvious which party was correct. We examined some of these disagreements and found a mixture of problems. In some cases, the clinicians did not follow the classification definitions, in others the classification rules seemed inferior to the actual clinical assessments; at other times, the computerized method was in error.

This study was not designed to measure how well 1 reader or method could predict neonatal outcome. Such an investigation would require much larger case numbers in each of the outcome categories and a different study design. An example of such an analysis is available.[20]

Generalizing these results to everyday practice must be done with caution. The results presented here reflect the combined effects of 5 distinct and unique factors: (1) a particular graded classification schema, (2) a specific set of tracings, (3) a select group of clinicians, (4) favorable working conditions that may be quite different from actual clinical practice, and (5) a specific set of computer algorithms. Changing any 1 of these factors could change the results.

The provision of quantitative information on the variation among clinicians and comparison with a computerized method that uses the same classification schema renders this study unique. We believe that agreement levels within 1 color code that are well over 80% are clinically acceptable and show that this rule-based approach provides a method for clinicians to achieve good consistency in EFM interpretation. Perfect agreement probably will never be achieved because each FHR segment requires the clinician to visually measure several factors, grade them, and then consider >134 possible combinations of them.

These findings have direct clinical relevance. Given that the degree of tracing abnormality determines the nature and urgency of intervention, it follows that consistent tracing evaluation is fundamental for clear communication and timely intervention with the appropriate measures. Therefore, it is particularly noteworthy that these clinicians were able to apply a complex 5-level classification consistently.

These observations do necessitate some comments about the strengths and weaknesses of both clinical- and computer-based methods for tracing evaluation and the reason that they should be viewed as complementary and not as 1 substituting for the other. In actual practice, clinicians are reasoning constantly and integrating myriads of information. For example, inferences about underlying pathophysiologic events, the observed response to interventions, and projections about the length of labor all affect their interpretation of the tracing. Clinicians can reason, and this computerized method cannot. Human judgment, however, will deteriorate under certain conditions (such as fatigue, crisis or where assessments are performed over long periods of time).[23] The psychologic phenomenon of "tunnel vision" refers to the tendency to perceive and confirm information that aligns with a particular viewpoint and to discard contradicting information. Variations of this include "framing bias," which refers to a tendency to create a coherent interpretation without examining all the available information, and "confirmation bias," which refers to seeking only the information that supports a particular opinion.[24] Computerization is impervious to these kinds of problems. It brings a consistent, objective analysis for consideration to counter these unhelpful conditions.

The use of EFM has evolved considerably since its inception. Standard nomenclature is now widespread. Methods for grading the tracing, based largely on clinical consensus and medical literature, are published and transformed to clinical policy.[8,13-18] These methods focus on classic EFM features that the human eye can recognize. With computerization, new options can be explored (such as applying statistical methods to determine what combinations of these features and/or what trends over time are the most discriminating). In addition, direct mathematic analysis of the tracing may identify aspects that are not apparent visually and will better discriminate the fetus who is at risk for hypoxic brain injury from normal fetuses.[25,26]

In the meantime, we must rely on existing methods for tracing evaluation. The analysis presented here has demonstrated that it is possible for these clinicians to apply 1 of the most complex classification methods with good interobserver agreement and that the computerized version performed in a fashion that is similar to this very select group of clinicians. Such software could be considered a way to provide an additional safety net; that is, the computer could be used as a decision-support tool in clinical obstetric care and in the education and training of obstetrics providers. ■

ACKNOWLEDGMENTS

We thank the 5 experts who so generously gave their time and expertise in the grading of approximately 90 hours of fetal heart rate tracings: Ernest Graham, MD, of Johns Hopkins University, Baltimore, MD; Tomoaki Ikeda, MD, PhD, National Cardiovascular Unit, Osaka, Japan; Tekoa King, CNM, MPH, University of California San Francisco, CA; Michael G. Ross, MD, MPH, University of California Los Angeles, CA; and Michael L. Socol, MD, Northwestern University, Chicago, IL. We also thank Alina Dyachenko, PhD, and Bruno Bendavid for statistical consulting and data analysis.

REFERENCES

1. Nielsen PV, Stigsby B, Nickelsen C, Nim J. Intra- and inter-observer variability in the assessment of intrapartum cardiotocograms. Acta Obstet Gynecol Scand 1987;66:421-4.

2. Beaulieu MD, Fabia J, Leduc B, et al. The reproducibility of intrapartum cardiotocogram assessments. Can Med Assoc J 1982;127:214-6.

3. 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;199:623.e1-5.

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

5. Devane D, Lalor J. Midwives' visual interpretation of intrapartum cardiotocographs: intra- and interobserver agreement. J Adv Nurs 2005;52:133-41.

6. Donker DK, van Geijn HP, Hasman A. Interobserver variation in the assessment of fetal heart rate recordings. Eur J Obstet Gynecol Reprod Biol 1993;52:21-8.

7. 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;183:361-6.

8. Clark SL, Belfort MA, Dildy GA, Meyers JA. Reducing obstetric litigation through alterations in practice patterns. Obstet Gynecol 2008;112:1279-83.

9. Joint Commission on Accreditation of Healthcare Organizations, USA. Preventing infant death and injury during delivery. Sentinel Event Alert 2004;30:1-3.

10. Ransom SB, Studdert DM, Dombrowski MP, Mello MM, Brennan TA. Reduced medicolegal risk by compliance with obstetric clinical pathways: a case-control study. Obstet Gynecol 2003;101:751-5.

11. Draper ES, Kurinczuk JJ, Lamming CR, Clarke M, James D, Field D. A confidential enquiry into cases of neonatal encephalopathy. Arch Dis Child Fetal Neonatal Ed 2002;87:F176-80.

12. Saphier CJ, Thomas EJ, Studdert D, Brennan TA, Acker D. Applying no-fault compensation criteria to obstetric malpractice claims. Prim Care Update Obstet Gynecol 1998;5:208-9.

13. National Institute of Child Health and Human Development Research Planning Workshop. Electronic fetal heart rate monitoring: research guidelines for interpretation. Am J Obstet Gynecol 1997;177:1385-90.

14. 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. Obstet Gynecol 2008;112:661-6.

15. American College of Obstetricians and Gynecologists. ACOG practice bulletin no. 106: intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles. Washington, DC: The College; 2009.

16. Association of Women's Health Obstetrics and Neonatal Nurses. Fetal heart monitoring: principles and practices. Lyndon A, L Ali, eds. Washington, DC: AWHON; 2009.

17. Royal College of Obstetricians and Gynecologists. Clinical effectiveness support unit. Evidence-based Clinical guideline no. 8: the use of electronic fetal monitoring. RCOG Press, London; 2001.

18. Liston R, Sawchuck D, Young D, et al. Fetal health surveillance: antepartum and intrapartum consensus guideline. J Obstet Gynaecol Can 2007;29(suppl):S3-56. [Published erratum appears in J Obstet Gynaecol Can 2007;29:909.]

19. Parer JT, Ikeda T. A framework for standardized management of intrapartum fetal heart rate patterns. Am J Obstet Gynecol 2007;197:26.e1-6.

20. Elliott C, Warrick PA, Graham E, Hamilton EF. Graded classification of fetal heart rate tracings: association with neonatal metabolic acidosis and neurologic morbidity. Am J Obstet Gynecol 2009;202:258.e1-8.

21. Homepage of the Fetal and Neonatal and Physiological Society. Available at: www.fnps-society.org/framework.pdf. Accessed: Jan. 11, 2010.

22. Lowry R. Kappa as a measure of concordance in categorical scoring in concepts and applications of inferential statistics. Available at: http://faculty.vassar.edu/lowry/webtext.html. Accessed Jan. 5, 2009.

23. Arnedt JT, Owens J, Crouch M, Stahl J, Carskadon MA. Neurobehavioral performance of residents after heavy night call vs after alcohol ingestion. JAMA 2005;294:1025-33.

24. Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 1998;2:175-200.

25. Warrick P, Hamilton E, Precup D, Kearney R. Identification of the dynamic relationship between intrapartum uterine pressure and fetal heart rate for normal and hypoxic fetuses. IEEE Trans Biomed Eng 2009;56:1587-97.

26. Warrick P, Hamilton E, Precup D, Kearney R. Classification of normal and hypoxic fetuses from systems modeling of intra-partum cardiotocography. IEEE Trans Biomed Eng In Press.