Description, evaluation and clinical decision making according to various fetal heart rate patterns

Inter-observer and regional variability

Øjvind Lidegaard, Lars Meinert Bøttcher and Tom Weber

From the Department of Obstetrics and Gynecology, Hvidovre Hospital, University of Copenhagen, Hvidovre, Denmark

Acta Obstet Gynecol Scand 1992; 71: 48–53

At 10 Danish obstetrical departments, 116 residents (42 senior and 74 junior) participated in a study to assess inter-observer and regional variability in the description and evaluation of and clinical decision regarding 11 fetal heart rate patterns. The 11 traces included normal as well as pathological patterns, and normal as well as clinically asphyxiated babies. Five antepartum and six intrapartum patterns were included. A total of 1,276 descriptions and evaluations were obtained. The degree of agreement in *description* of fetal heart rate changes was high regarding the baseline and the presence of silent or sinusoidal pattern (87–94% on an arbitrary 0–100% scale), and low regarding the assessment of variability and type of deceleration (50–72%). The degree of agreement in *interpreting* heart rate patterns was 59% (on an arbitrary 0–100% scale). Senior residents generally interpreted the changes as indicative of less serious fetal stress than did their junior colleagues, explaining why junior residents 30% more frequently than their older colleagues found an indication for Cesarean section. Relatively low regional inter-observer agreement scores were primarily due to low agreement between departments, especially between departments far apart. It is concluded that we still need a scientific clarification of which specific heart rate changes are the best predictors of fetal stress. Artificial intelligence programs for interpreting fetal cardiotocograms and ECG signals constitute one promising prospect.

Key words: fetal heart rate evaluation; electronic fetal monitoring; inter-observer variability; cardiotocogram evaluation; regional variation

Submitted April 4, 1991
Accepted September 18, 1991

During recent decades, several new techniques have been introduced into the management of labor with the aim of providing maximum safety for fetus and mother. Unfortunately, a number of these techniques have been introduced without preceding technology assessment, leaving us without knowledge of their exact benefits and risks. The main technique in obstetrics is electronic fetal heart rate monitoring, the efficacy of which has been the subject of several studies (1–5). Most studies have found some benefit in high-risk deliveries, while its benefit in routine use is still controversial.

Years after the introduction of new techniques, technological assessment is often difficult and sometimes impracticable. The study of regional variation in medical practice, however, may be one possible approach to evaluating medical techniques or routines; the documentation of major regional and/or inter-observer variability being indicative of a non-consistent scientific basis and/or insufficient clinical education (6, 7).

The relation between fetal heart rate changes and the clinical and metabolic condition of the newborn has been analysed on previous occasions (8–10). The varying correlations found may be explained by the fact that the majority of heart rate changes merely indicate an increased probability of fetal stress. As even a slightly increased risk of severe fetal stress justifies intervention, despite the fact that most fetuses exhibiting pathological changes are quite normal, studies to assess the sensitivity and specificity of fetal heart rate changes for fetal asphyxia require at least a large number of monitored deliveries and a clinical team of fixed composition. This circumstance may explain the limited number of conclusive studies undertaken so far.

The purpose of this study was to assess inter-observer and regional variability in the description and evaluation of fetal heart rate patterns among Danish obstetrical residents who interpret these traces daily and the significance of these differences for their clinical decisions.

Material and methods

Ten obstetrical departments where 36% of all births taking place in Denmark each year were involved in the study. The departments represented all three major regions in Denmark (Sjælland, Fyn and Jylland) and different sizes of departments. The head of each department was forwarded eleven fetal heart rate patterns and corresponding clinical data. During the same week in October 1989 all residents on duty were presented with the 11 cases for the first time and were asked to describe and evaluate the heart rate patterns individually. They were given all relevant clinical information and were asked for their clinical decision regarding each case (± indication of Cesarean section). The description was noted on a precoded questionnaire, specifying different changes. The interpretation of these descriptions was graded on an arbitrary ordinal scale indicating a non, slightly, moderately or severely stressed fetus.

The eleven patterns were chosen so as to include both normal and pathological traces and traces from deliveries with normal and asphyctic babies. This information was given to the participating residents.

Statistics

To compare the inter-observer variability in describing different fetal heart rate changes, an agreement score was calculated as

$$\left(\left(\frac{n_\mathrm{a}}{N_\mathrm{a}} + \frac{n_\mathrm{b}}{N_\mathrm{b}} + \ldots + \frac{n_k}{N_\mathrm{k}} \right) /11 \times 100\% - 50\% \right) \times 2,$$

where n is the number of answers given by the majority at each trace, and N the total number of residents describing the pattern. According to this arbitrary scale, 100% indicates total agreement, 0% total disagreement.

In the evaluation (interpretation) of the eleven patterns, an agreement score (e.g. according to each department) for the eleven heart rate patterns was calculated as:

$$\left(\left(\frac{n_\mathrm{a} - \mathrm{o}_\mathrm{a} - 2 \times p_\mathrm{a}}{N_\mathrm{a}} + \frac{n_\mathrm{b} - \mathrm{o}_\mathrm{b} - 2 \times p_\mathrm{b}}{N_\mathrm{b}} + \ldots + \frac{n_\mathrm{k} - \mathrm{o}_\mathrm{k} - 2 \times p_\mathrm{k}}{N_\mathrm{k}} \right) /11 \times 100\% + 50\% \right) \times 2/3$$

where

n is the number of persons who, of four ordinal answers (non, slightly, moderately or severely stressed fetus), chose the response which most had chosen,

o is the number of persons choosing an answer which was 2 steps from the 'majority answer',

p is the number of persons choosing an answer which was 3 steps (if possible) from the majority answer.

N is the total number of residents evaluating that particular heart rate pattern.

According to this arbitrary 0–100 point scale, 16 points correspond a random distribution. When calculating an average for more departments or regions, all responders at each pattern were inserted for n, o, p, and N, implying a weighting of the different departments according to their respective number of responders.

The large sample behaviour of the two agreement scores described above was asymptotically normal. Means and standard deviations therefore had to be calculated separately for each subgroup of responders, assuming independent evaluations of separate patterns.

Table I. Agreement in description and evaluation of specific FHM patterns and in clinical decision among 116 Danish residents; degree of agreement is indicated by an arbitrary agreement score (0–100%)

	Agreement score	(SD)
Description		
Normal pattern	79.9%	1.6
Bradycardia	89.1%	1.2
Tachycardia	91.8%	1.1
Reduced variability	52.2%	2.2
Silent pattern	86.7%	1.2
Early decelerations	71.7%	1.7
Variable decelerations	49.7%	2.4
Late decelerations	54.5%	2.2
Sinusoidal pattern	94.0%	0.9
Evaluation[1]		
Range (for different FHM patterns)	26–80%	
Mean	59.3%	1.3
Junior residents:	59.3%	1.6
Senior residents:	63.9%	2.0
Clinical decision[2]		
Range (for different FHM patterns)	53–99%	
Mean	69.5%	1.8

[1] non-stressed, slightly stressed, moderately stressed or severely stressed fetus.

[2] ± Cesarean section.

Results

The total number of participating residents was 116 (74 junior, 42 senior), each evaluating 11 fetal heart rate patterns, giving 1,276 descriptions and evaluations. Of ten participating departments, three had less than 1,500 births per year, four between 1,500 and 2,500 and three more than 2,500 births per year.

Description

The description of the fetal heart rate patterns is shown in Table I. The inter-observer variability, indicated by an agreement score, varied for different heart rate changes between 49.7% and 94.0%. Most controversial were the presence of reduced variability, variable decelerations and late decelerations.

There were no significant differences in the description of fetal heart rate patterns according to the charge of the residents, major region, or size of department.

Evaluation

The evaluation of the fetal heart rate changes on a 4-step ordinal scale is displayed in Table I. The inter-observer variability in this interpretation was 59.3% overall (according to the above-mentioned agreement score), ranging between 25.7% and 79.7% for different patterns.

Senior residents generally interpreted the heart rate changes as indicative of less serious fetal stress, than did their junior colleagues. On average, 52.8% of the junior residents found the changes to be indicative of moderate or severe fetal stress, versus 45.9% of senior residents ($p < 0.01$). In 7 of the 11 patterns, junior residents interpreted the traces as indicative of more severe stress, than did their senior colleagues. Furthermore, senior residents obtained a better agreement score (63.7%) than junior residents (59.3%) ($p = 0.08$) (Table I).

The relationship between the type of change described and the interpretation of these changes is

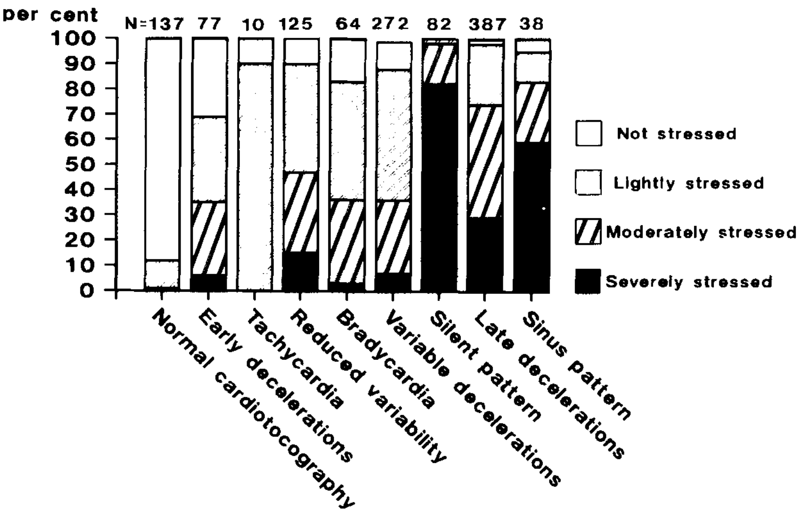

Fig. 1. Evaluation of specific fetal heart rate patterns by 116 Danish residents. When more than one change was present, that pattern was allocated to the most severe change according to the ordinal succession in the Figure. Thus *n* represents the number of residents signifying that specific pattern and not more severe changes.

Table II. Relationship between evaluation of fetal heart rate patterns and clinical decision (± Cesarean section (CS) according to charge of residents

Charge of residents:	Junior residents	Senior residents	All residents
Number of completed answers	750	442	1192
Fetal heart rate evaluation	+CS	+CS	+CS
Non-stressed fetus	2.5%	0.0%	1.4%
Slightly stressed fetus	1.3%	1.3%	1.3%
Moderately stressed fetus	17.7%	15.0%	16.8%
Severely stressed fetus	87.8%	72.2%	82.1%
Average	24.8%	19.0%	22.7%

illustrated in Fig. 1. Changes which had a low descriptive agreement score were also difficult to interpret. The change which was found indicative of the most severe fetal stress was silent pattern, followed by sinusoidal pattern and late decelerations.

The degree of agreement in the evaluation of the different patterns at the 10 participating centres ranged between 55.5% and 69.1% (p NS).

The degree of agreement in the evaluation of the 11 cases among the three regions included ranged between 56.9% and 62.5% (p NS). In this calculation, all participants in each region were analysed together. Alternatively if the region score was calculated as the average of the scores at each department in that region (weighted according to the number of participants), the range was narrowed.

Finally, the degree of agreement in evaluation according to size of department ranged between 56.6% and 64.5% (p NS), the middle-sized departments having the highest scores.

These differences according to region and size of department were not attributable to differences in the distribution of senior and junior residents.

Clinical decision

For each fetal heart rate pattern and corresponding clinical information, the participants signified whether they found an indication for Cesarean section (CS) (Table I). Analysing all 11 traces together (Table II), 24.8% of junior residents found an indication for CS, vs. 19% of senior residents ($p <$ 0.05). Thus junior residents found an indication for CS 30% more often than did the senior residents.

An analysis of the percentage of senior and junior residents who found indication for CS revealed that their mutual differences were due primarily to the fact that senior residents interpreted the patterns as indicative of less serious stress than did junior residents (Table II), while only minor differences were found in description of the heart rate patterns and in indications for CS awarded a certain value.

Discussion

The number of fetal heart rate patterns included in this study does not allow analysis of the agreement between the description and evaluation of these patterns and the fetal outcome, as such a correlation would be entirely dependent on the selection of the 11 cases. Although the heart rate patterns included in this study all represented traces with which our Department had been concerned, heart rate patterns giving rise to concern are of primary interest, because they influence our clinical decisions. Therefore, the degree of agreement in description and evaluation of the heart rate patterns included in this study were *a priori* expected to be significantly less than would have been the case, had the agreement been based on a representative sample of heart rate traces, which would comprise mainly normal patterns.

Generally speaking, the agreement in the description of the heart rate patterns was better than the agreement in evaluation. This difference, however, reflects primarily the fact that the description was dichotomous whereas the evaluation was on an ordinal four-step scale. Consequently the descriptive and evaluative agreement scores are not directly comparable.

Table I, however, reveals that this overall fairly high descriptive agreement of heart rate changes conceals a rather wide variation as regards the type of abnormality, especially concerning the presence of reduced variability and type of deceleration. These differences in the detailed description partly account for the low agreement scores in the interpretation of the heart rate patterns. As clinical decisions are based upon interpretations, it is primarily the degree of agreement in the evaluation rather than in the description which is of concern.

The variability in the agreement in the evaluation between different heart rate patterns reflects patterns of different degrees of interpretation difficulties. Major variations in the interpretation of medical data are not exceptional. Of particular interest is the demonstration of dispersion in the description of electrocardiograms (ECG) (11, 12) and ultrasound examinations (13), not much different from the present average agreement score of 59.3%. As long as we are dealing with evaluations which have a major element of descriptive interpretation, variations in this size are probably more the rule than the exception.

In 1982, Lotgering et al. investigated inter-observer and intra-observer variation in the description of 100 cardiograms among five experienced observers, and found K-values between 9% and 69% (on a − 100% to + 100% scale) (14). Thus, our interpretations in 1989 have not improved substantially compared with those in The Netherlands in 1982.

Training seems to be of some significance, as senior residents obtained better evaluation agreement scores than junior residents (63.9% vs. 59.3%). Although all heart rate patterns included in this study were selected as traces which had given rise to concern, an agreement score of about 60% partly reflects the lack of valid knowledge as to which specific heart rate patterns indicate fetal stress. This supposition is supported by the dispersion in clinical evaluation for specific heart rate changes (Fig. 1). In Denmark, the lack of systematic training concerning fetal heart rate interpretations may also be of significance.

Of further interest is the fact that a silent pattern was interpreted as indicative of more severe fetal stress than were sinusoidal pattern and late decelerations, which is not in accordance with the consensus of opinion. A valid scientific clarification as to which heart rate changes best predict fetal stress could provide a basis for a more systematic clinical training and education. The development of intelligent programs which analyse fetal heart rate patterns in a reproducible way seems to be one possible way to prospectively indentify those changes which have the highest predictive value (15, 16). In the long term, the development of artificial intelligence programs for analysing fetal heart rate patterns as well as fetal ECG signals offers promising prospects (17, 18).

The clinical decisions were based on evaluation of fetal heart rate changes, as well as on the clinical data presented. The overall agreement score of 70% suggests that the knowledge of clinical data improves the inter-observer variability only slightly (from 60% to 70%). On the other hand, there was no obvious correlation between a high agreement score in the evaluation of each fetal heart rate pattern and the clinical agreement in the same case, suggesting that the clinical data were important for the decision to undertake a CS. In Denmark the CS rate on different maternity wards ranges between 6% and 25%. On average, two-thirds of these are non-elective sections. As long as our interpretation of available clinical data and fetal heart rate changes is characterized by inter-observer agreement scores of about 50–70% we cannot expect a much narrower regional variability in CS rates.

Conclusion

The degree of agreement in the description of fetal heart rate changes was high as regards the baseline and the presence of silent pattern and sinusoidal pattern, but low concerning the assessment of variability and kind of deceleration.

Older residents generally interpreted the heart rate changes as indicative of less serious fetal stress than did their younger colleagues, which explains why junior residents found an indication for CS 30% more often than their senior colleagues. The difference in inter-observer variability in evaluation between three regions in Denmark and according to the size of departments did not reach significance level.

It is concluded that we still need a scientific clarification of which specific heart rate changes constitute the best predictors of fetal stress. Artificial intelligence programs for interpreting fetal CTG and ECG signals offer one promising prospect.

Acknowledgement

We are indebted to Svend Kreiner, statistician, for his statistical analyses.

References

1. Haverkamp AD, Orleans M, Langendoerfer S, et al. A controlled trial of the differential effects of intrapartum fetal monitoring. Am J Obstet Gynecol 1979; 134: 399–408.
2. Greenland S, Olsen J, Rachootin P, Pedersen GT. Effects of electronic fetal monitoring on rates of early neonatal death, low Apgar score, and Cesarean section. Acta Obstet Gynecol Scand 1985; 64: 75–80.
3. MacDonald D, Grant A, Sheridan-Pereira M, et al. The Dublin randomized controlled trial of intrapartum fetal heart rate monitoring. Am J Obstet Gynecol 1985; 152: 524–39.
4. Leveno KJ, Cunningham FG, Nelson S, et al. A prospective comparison of selective and universal electronic fetal monitoring in 34,995 pregnancies. N Engl J Med 1986; 315: 615–9.
5. Ingemarsson I, Arulkumaran A, Ingemarsson E, Tambyraja RL, Ratnam SS. Admission test: A screening test for fetal distress in labor. Obstet Gynecol 1986; 68: 800–6.
6. American Medical Association. Confronting regional variations. The Maine approach 1986.
7. Rothberg DL. Regional variations in hospital use, geographic and temporal patterns of care in the United States. University Health Policy Consortium Series. Lexington, 1982.
8. Montan S, Olofsson P, Solum T. Classification of the

nonstress test and fetal outcome in 1056 pregnancies. Acta Obstet Gynecol Scand 1985; 64: 639–44.

9. 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.
10. Wennergren M, Krantz M, Hjalmarson O. Fetal heart rate pattern and risk for respiratory disturbance in full-term newborns. Obstet Gynecol 1986; 68: 49–53.
11. Gorman PA, Calatayud JB, Abraham S, Caceres CA. Observer variation in interpretation of electrocardiogram. Med Ann Distr Columbia 1964; 33: 97–9.
12. Davis LG. Observer variation in reports on electrocardiograms. Br Heart J 1958; 20: 153–61.
13. Sarmandal P, Bailey SM, Grant JM. A comparison of three methods of assessing inter-observer variation applied to ultrasonic fetal measurement in the third trimester. Br J Obstet Gynaecol 1989; 96: 1261–5.
14. Lotgering FK, Wallenburg HCS, Schouten HJA. Interobserver and intraobserver variation in the assessment of antepartum cardiotocograms. Am J Obstet Gynecol 1982; 144: 701–5.
15. Stigsby B, Nielsen PV, Nickelsen C, Nim J. Computer assessment of the intrapartum cardiotocograms. Methods of data reduction and diagnostic procedure. Acta Obstet Gynecol Scand 1988; 67: 455–60.
16. Nielsen PV, Stigsby B, Nickelsen C, Nim J. Computer assessment of the intrapartum cardiotocogram. The value of computer assessment compared with visual assessment. Acta Obstet Gynecol Scand 1988; 67: 461–4.
17. Murray HG. The fetal electrocardiogram: current clinical developments in Nottingham. J Perinat Med 1986; 14: 399–404.
18. Rosén KG. Alterations in the fetal electrocardiogram as a sign of fetal asphyxia. Experimental data with a clinical implementation. J Perinat Med 1986; 14: 355–63.

Address for correspondence:

Øjvind Lidegaard, M.D.
Olgasvej 23
DK-2950 Vedbaek
Denmark