International Journal of Gynecology & Obstetrics 57 (1997) 33–37

International Journal of GYNECOLOGY & OBSTETRICS

Article

Evaluation of interobserver agreement of cardiotocograms

J. Bernardes[a]*, A. Costa-Pereira[b], D. Ayres-de-Campos[a], H.P. van Geijn[c], L. Pereira-Leite[a]

[a] *Departamento de Ginecologia e Obstetrícia, Hospital de S. João, Faculdade de Medicina do Porto, Oporto, Portugal*
[b] *Serviço de Bioestatística e Informática Médica, Faculdade de Medicina do Porto, Oporto, Portugal*
[c] *Free University Hospital, Department of Obstetrics and Gynaecology, Amsterdam, The Netherlands*

Received 18 October 1996; accepted 8 January 1997

Abstract

Objective: To evaluate interobserver agreement in visual analysis of each cardiotocographic event. *Methods:* Three experts independently divided 16 antepartum and 17 intrapartum cardiotocograms into baseline segments, accelerations and decelerations, according to the FIGO guidelines. Baseline segments were further classified as having normal, reduced or increased variability and decelerations as early, late and variable. Uterine activity was divided into tonus and contractions. Agreement was assessed by the proportions of agreement (*pa*) with 95% confidence intervals. *Results:* Reproducibility in assessment of baseline segments with normal variability, accelerations and uterine activity was acceptable (*pa* = 0.56–0.71) whereas that of other segments was not (*pa* = 0.14–0.45). *Conclusions:* Analysis of most cardiotocographic events is poorly reproducible, even when experts use the FIGO guidelines. This may be explained by some still ambiguous guidelines, by eyeball limitations in evaluation of subtle events, and by the incapacity of busy clinicians to assess complex and multiple cardiotocographic events in a systematic and disciplined fashion. © 1997 International Federation of Gynecology and Obstetrics

Keywords: Fetal heart rate; Reproducibility; Computerized analysis

1. Introduction

Many authors have reported on the poor reproducibility of visual analysis of cardiotocograms and on the negative clinical and medico-legal consequences of this finding [1–5]. However, the reasons for these observations are not yet clear. This may arise because various criteria for cardiotocogram analysis were employed. Different observer expertise may also have been a con-

* Corresponding author. Tel.: +351 2 527151 (ext. 1140); fax: +351 2 5505870.

0020-7292/97/$17.00
PII S0020-7292(97)02846-4

tributing factor. Moreover, limitations in the statistical analysis of the previous studies related to this subject have been claimed [6,7]. These statistical limitations may have provided biased overall agreement results in cardiotocogram analysis, rather than unbiased agreement results by each cardiotocographic event, as recognized by some of the authors of the mentioned studies [6] and as demonstrated by Grant [7].

Recently, we have shown [8] that a more objective and reproducible fetal heart rate (FHR) baseline could be estimated using the International Federation of Gynecology and Obstetrics (FIGO) guidelines for fetal monitoring [9]. Using the same cardiotocograms, in this paper we evaluate agreement in analysis of each of the cardiotocographic events, as defined by the FIGO guidelines: baseline, accelerations, decelerations, variability and uterine activity. It was hoped that this could shed some light on the reason why overall analysis of cardiotocograms was poorly reproducible. The proportions of agreement, claimed as the only appropriate method to assess interobserver variation of categorical variables [7], were used. The kappa statistic was also employed, for comparison with other results.

2. Materials and methods

Three obstetricians with acknowledged expertise in fetal monitoring, working in different teaching hospitals, were asked to analyze 33 cardiotocograms following the FIGO guidelines for fetal monitoring. Sixteen antepartum and 17 intrapartum cardiotocograms, originating from 22 third-trimester high-risk pregnancies, were randomly selected. In 8 cases, recordings started in the first stage of labor and continued throughout the second stage. Tracings were recorded by a Toitu MT 810-B fetal monitor using echo-Doppler with autocorrelation in the antepartum and the fetal electrocardiogram in the intrapartum. Uterine activity was assessed by tocodynamometry and fetal movements were registered as perceived by the mother. Paper speed was 1 cm/min. Data on the duration of tracings and gestational age on monitoring are given in Table 1.

Table 1
Number (n), mean duration (±S.D.) and mean gestational age (±S.D.) corresponding to the ante and intrapartum cardiotocograms included in the study

Cardiotocograms	Antepartum	Intrapartum	
		1st stage	1st and 2nd stage
n	16	9	8
Duration (min)	41 (±14)	64 (±21)	51 (±31)
Gestational age (weeks)	34 (±4)	40 (±1)	40 (±1)

Tracings were sent to experts by mail, to be returned within 3 months. The FIGO guidelines and all relevant clinical information on the cases, prior to cardiotocogram recording were provided. The latter included gestational age, administered drugs, body temperature and stage of labor. Instructions to divide tracings with vertical bars, into FHR baseline segments, accelerations, decelerations, uterine contractions and uterine activity baseline were enclosed. Experts were further asked to classify baseline segments as having normal, reduced or increased long-term variability and decelerations as being early, variable or late. FHR and uterine activity segments not corresponding to any of the previous categories were to be classified as 'other segments'. A case-example was included.

Only segments clearly identified by the three observers were included in the study. Segments with 2 mm or less were ignored. Contiguous accelerations, i.e. without baseline in between, were considered as one single event. The same was decided for decelerations.

Agreement was assessed by the proportions of agreement with 95% confidence intervals (CI) as described by Grant [7]. According to this author, the proportion of agreement that signifies good agreement is arbitrary, but if the 95% CI includes 0.50, then agreement is almost certainly poor (provided the study population is large enough). For each segment, three trials of agreement between observers A, B and C were analyzed (A with B; B with C; A with C). Agreement beyond chance was also assessed, using the κ statistic.

Kappa values larger than 0.75 were considered indicators of excellent agreement, those between 0.40 and 0.75 as indicators of fair to good agreement, and those below 0.40, as indicators of poor agreement [3,7]. Agreement was first assessed regarding detection of FHR baseline segments, accelerations, decelerations, other FHR segments, uterine contractions, tonus and other uterine activity segments (Table 2). Then, in segments unanimously considered FHR baseline, agreement in classification of long-term variability as normal, decreased or increased was assessed. Likewise, in segments unanimously considered decelerations, agreement was evaluated in their classification as early, late or variable (Table 3).

3. Results

The total number of agreement trials between the three observers was 6535. The number of agreement trials in each assessed category (e.g. accelerations or decelerations) is not always a multiple of three because frequently the same segment was classified in different categories by different observers.

Table 2
Agreement regarding visual detection of ante and intrapartum FHR and uterine activity segments. Number (n) of agreement trials for each segment category, proportions of agreement (pa) with 95% confidence intervals (95% CI) and κ statistic

	n	pa	95% CI	κ statistic
Antepartum				
FHR				
Baseline	574	0.63	0.59–0.67	0.47
Accelerations	319	0.57	0.52–0.62	0.53
Decelerations	116	0.26	0.18–0.34	0.26
Others	87	0.18	0.10–0.26	0.18
Uterine activity				
Contractions	144	0.72	0.65–0.79	0.70
Tonus	332	0.70	0.65–0.75	0.51
Others	96	0.07	0.00–0.18	0.06
Intrapartum				
FHR				
Baseline	872	0.63	0.60–0.66	0.51
Accelerations	507	0.56	0.52–0.60	0.52
Decelerations	332	0.51	0.46–0.56	0.49
Others	130	0.22	0.15–0.29	0.22
Uterine activity				
Contractions	828	0.70	0.67–0.73	0.59
Tonus	604	0.62	0.57–0.67	0.56
Others	224	0.10	0.05–0.15	0.08
Overall				
FHR				
Baseline	1446	0.63	0.61–0.65	0.49
Accelerations	826	0.56	0.53–0.56	0.53
Decelerations	448	0.45	0.40–0.50	0.43
Others	217	0.21	0.16–0.26	0.21
Uterine activity				
Contractions	972	0.71	0.67–0.73	0.62
Tonus	936	0.62	0.57–0.67	0.56
Others	320	0.10	0.05–0.15	0.08

Table 3
Agreement regarding visual detection of ante and intrapartum long-term variability and deceleration's classification. Number (n) of agreement trials for each segment category, proportions of agreement (pa) with 95% confidence intervals (95% CI) and κ statistic

	n	pa	95% CI	κ statistic
Antepartum				
Long-term variability				
Normal	256	0.69	0.62–0.74	0.41
Reduced	94	0.53	0.43–0.63	0.51
Increased	47	0.15	0.05–0.25	0.14
Deceleration's classification				
Variable	0	—	—	—
Early	12	0.67	0.40–0.94	0.53
Late	10	0.60	0.30–0.90	0.51
Intrapartum				
Long-term variability				
Normal	465	0.64	0.60–0.68	0.34
Reduced	235	0.40	0.34–0.46	0.35
Increased	30	0.13	0.04–0.31	0.13
Deceleration's classification				
Variable	107	0.27	0.19–0.35	0.05
Early	42	0.31	0.20–0.42	0.23
Late	72	0.24	0.11–0.37	0.21
Overall				
Long-term variability				
Normal	721	0.66	0.63–0.69	0.37
Reduced	329	0.44	0.39–0.49	0.40
Increased	77	0.14	0.06–0.22	0.14
Deceleration's classification				
Variable	107	0.27	0.19–0.35	0.03
Early	84	0.36	0.26–0.46	0.15
Late	52	0.31	0.18–0.44	0.32

Interobserver agreement in the detection of baseline segments, accelerations and uterine contractions was fair to good with overall proportions of agreement and κ statistics ranging from 0.56–0.71 and 0.49–0.62, respectively (Table 2). Agreement in detection of decelerations was poor with an overall proportion of agreement and κ statistic of 0.45 and 0.43, respectively (Table 2). In 30 agreement trials the same FHR segment was classified as an acceleration and a deceleration by two different referees.

In the classification of long-term variability and decelerations, the total number of agreement trials was 1370. As shown in Table 3, agreement in classification of long-term variability was only acceptable when it was considered normal (pa = 0.66). For abnormal variability and classification of decelerations, agreement was poor (pa = 0.14–0.44 and κ = 0.03–0.40).

Agreement on deceleration detection and classification as early, variable or late was significantly worse in the intrapartum than in the antepartum.

4. Discussion

In order to minimize conditions for disagreement, cardiotocogram analysis in our study was performed by experienced clinicians, according to the FIGO guidelines. Long tracings were analyzed with uniform criteria, in the context of the clinical situation. A paper speed of 1 cm/min was used for recording. This is the routine procedure in our department as in many other centers throughout the world. It may be argued that it may have negatively influenced results, however poor reproducibility has also been demonstrated with paper speeds of 3 cm/min [10].

The proportions of agreement are claimed as the most appropriate statistical method for measuring interobserver agreement. The kappa statistic, which is frequently used for this purpose, appears to have several drawbacks. It tests whether the association of assessments by observers is due to chance, but has nothing to do with their agreement [7]. Taking our results as an example, when FHR variability was classified as normal with a prevalence as high as 64%, the proportion of agreement was 0.66 (95% CI 0.63–0.69) whereas the κ statistic was only 0.37 (Table 3).

Our results suggest that analysis of most cardiotocographic events, even when performed by experienced clinicians, with uniform criteria and access to clinical information, as recommended in the FIGO guidelines, is poorly reproducible. They also suggest that reproducibility in the detection of the more gross and stable cardiotocogram events (FHR baseline segments with normal variability, accelerations and uterine contractions) is acceptable. Conversely, detection of subtle alterations (different kinds of decelerations, reduced or increased FHR variability) is only poorly reproducible.

We propose three main reasons for these results. Firstly, ambiguous definitions of cardiotocographic events are still present even in the FIGO guidelines. For example, classification of decelerations as early, variable or late is never precisely defined. Also there is an ambiguous interdependence of definitions of FHR baseline, accelerations and decelerations: baseline is defined as the mean FHR in the absence of accelerations and decelerations and the latter are defined as transient changes in FHR in respect to the baseline [8]. Secondly, a reproducible eyeball evaluation of very subtle cardiotocographic alterations, such as decreased long-term variability, might be difficult or even impossible. Thirdly, a systematic and disciplined assessment of the cardiotocogram's complex and multiple events by busy clinicians may also be impossible. Such is, for instance, the case of the evaluation of long-term variability throughout the whole tracing.

Revision of guidelines for fetal monitoring such as those of the FIGO to include a more precise definition of cardiotocographic events could probably to some extent improve agreement. However, human visual and methodological inaccuracy would still remain a problem. A more disciplined method of cardiotocogram analysis e.g. prolonged examination and the systematic use of a ruler, could probably improve agreement, but this is usually not possible in busy clinical prac-

tice. Both visual and methodological inaccuracies can probably only be reliably overcome by computerized analysis.

In conclusion, this study suggests that clinicians may be reasonably confident regarding the reproducibility of their findings when they detect FHR baseline segments with normal long-term variability or accelerations. Moreover, reproducibility of visual detection of uterine contractions, even when obtained by tocodynamometry, is acceptable both in ante and intrapartum tracings. However, they should be extremely cautious when identifying decelerations or FHR baseline segments with reduced or increased long-term variability. In this case, it may help to review or prolong the tracing and/or even call for a second opinion. A more disciplined detection of events with the use of a ruler, should also be considered. However, an entirely reproducible analysis of cardiotocographic events can only be accomplished by computerized analysis.

In the last decade many studies including randomized controlled trials have embodied a controversy about the uncertain clinical value of cardiotocography [11–15]. We believe that our study is a step forward to overcome this apparently endless controversy, namely by providing objective insights for future research in visual and/or computerized cardiotocographic analysis, including randomized controlled trials.

Acknowledgments

Professors Luis Graça, Paulo Moura and Santos Jorge are kindly acknowledged for their collaboration in analysis of cardiotocograms. This study was supported by grant 28757, Instituto Nacional de Investigação Científica, Portugal, by grant 87 160/MIC, Instituto Nacional de Investigação Científica e Tecnológica, Portugal and by EC Concerted Action 'New Methods for Perinatal Surveillance' project no. II.1.1/7.

References

[1] Mohide P, Keirse MJNC. Biophysical assessment of fetal well-being.In: Chalmers I, Enkin M, Keirse MJNC, editors, Effective Care in Pregnancy and Childbirth. Oxford: University Press. 1989: 477.

[2] Grant A. Monitoring the fetus during labour. In: Chalmers I, Enkin M, Keirse MJNC, editors, Effective Care in Pregnancy and Childbirth. Oxford: University Press. 1989: 846.

[3] 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.

[4] Boehm FH. Fetal distress. In: Eden RE, Boehm FH, editors, Assessment and Care of the Fetus; Physiological, Clinical and Medico-legal Principles. East Norwalk: Prentice-Hall International Inc. 1990: 809.

[5] Symonds EM. Litigation and the intrapartum cardiotocogram. Br J Obstet Gynaecol 1993; 100 (Suppl 9): 8.

[6] Donker DK, Hasman A, Van Geijn HP. Interpretation of low kappa values.Int J Biomed Comput 1993; 33: 55.

[7] Grant JM. The fetal heart rate is normal, isn't it? Observer agreement of categorical assessments. Lancet 1991; 337: 215.

[8] Bernardes J, Costa-Pereira A, van Geijn HP, Pereira-Leite L. A more objective fetal heart rate baseline estimation. Br J Obstet Gynaecol 1996; 103: 714.

[9] Rooth G, Huch A, Huch R: Guidelines for the use of fetal monitoring. Int J Gynecol Obstet 1987; 25: 159.

[10] Lotgering FK, Wallenburg HCS, Schouten HJA. Interobserver and intraobserver variation in the assessment of antepartum cardiotocograms. Am J Obstet Gynecol 1982; 144: 701.

[11] Nelson KB, Dambrosia JM, Ting TY, Grether JK. Uncertain value of electronic fetal monitoring in predicting cerebral palsy. N Engl J Med 1996; 334: 613.

[12] MacDonald D. Cerebral palsy and intrapartum fetal monitoring. N Engl J Med 1996; 334: 659.

[13] Thacker SB, Stroup DF, Peterson HB. Efficacy and safety of intrapartum electronic fetal monitoring: an update. Obstet Gynecol 1995; 86: 613.

[14] Bernardes J, Costa-Pereira A. Efficacy and safety of intrapartum electronic fetal monitoring (letter). Obstet Gynecol 1996; 87: 476.

[15] Parer JT. Efficacy and safety of intrapartum electronic fetal monitoring (letter). Obstet Gynecol 1996; 87: 476.