Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

Agreement on intrapartum cardiotocogram recordings between expert obstetricians

Lukáš Hruban MD,[1] Jiří Spilka PhD,[2,4] Václav Chudáček PhD,[2,4] Petr Janků MD PhD,[1] Michal Huptych PhD,[2] Miroslav Burša PhD,[2] Adam Hudec MD,[5] Marian Kacerovský MD PhD,[6] Michal Koucký MD PhD,[7] Martin Procházka MD PhD,[8] Vladimír Korečko MD,[5] Jan Seget'a MD,[9] Ondřej Šimetka MD PhD,[9,10] Alena Měchurová MD PhD[11] and Lenka Lhotská PhD[3]

[1]Consultant obstetrician, Department of Gynecology and Obstetrics, Masaryk University Hospital, Brno, Czech Republic
[2]Researcher, [3]Associate professor, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[4]Researcher, Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Prague, Czech Republic
[5]Consultant obstetrician, Department of Gynecology and Obstetrics, University Hospital in Plzeň, Plzeň, Czech Republic
[6]Consultant obstetrician, Department of Gynecology and Obstetrics, University Hospital in Hradec Králové, Hradec Králové, Czech Republic
[7]Consultant obstetrician, Department of Gynecology and Obstetrics, University Hospital in Prague, Prague, Czech Republic
[8]Associate professor, Department of Gynecology and Obstetrics, University Hospital in Olomouc, Olomouc, Czech Republic
[9]Consultant obstetrician, Department of Gynecology and Obstetrics, University Hospital Ostrava, Ostrava, Czech Republic
[10]Associate professor, Department of Surgical Studies, Ostrava University, Ostrava, Czech Republic
[11]Associate professor, Department for Mother and Child Care, Prague Podolí, Prague, Czech Republic

Keywords
cardiotocography, fetal heart rate, inter-observer variability, intra-observer variability

Correspondence
Dr Jiří Spilka
Czech Institute of Informatics, Robotics and Cybernetics
Czech Technical University in Prague
Prague 6, 166 36
Czech Republic
E-mail: jiri.spilka@ciirc.cvut.cz

Accepted for publication: 10 March 2015

doi:10.1111/jep.12368

Abstract

Rationale, aims and objectives To evaluate obstetricians' inter- and intra-observer agreement on intrapartum cardiotocogram (CTG) recordings and to examine obstetricians' evaluations with respect to umbilical artery pH and base deficit.

Methods Nine experienced obstetricians annotated 634 intrapartum CTG recordings. The evaluation of each recording was divided into four steps: evaluation of two 30-minute windows in the first stage of labour, evaluation of one window in the second stage of labour and labour outcome prediction. The complete set of evaluations used for this experiment is available online. The inter- and intra-observer agreement was evaluated using proportion of agreement and kappa coefficient. Clinicians' sensitivity and specificity was computed with respect to umbilical artery pH, base deficit and to Apgar score at the fifth minute.

Results The overall proportion of agreement between clinicians reached 48% with 95% confidence intervals (CI) (CI: 47–50). Regarding the different classes, proportion of agreement ranged from 57% (CI: 54–60) for normal to 41% (CI: 36–46) for pathological class. The sensitivity of clinicians' majority vote to objective outcome was 39% (CI: 16–63) for the umbilical artery base deficit and 27% (CI: 16–42) for pH. The specificity was 89% (CI: 86–92) for both types of objective outcome.

Conclusions The reported inter-/intra-observer variability is large and this holds irrespective of clinicians' experience or work place. The results support the need of modernized guidelines for CTG evaluation and/or objectivization and repeatability by introduction of a computerized approach that could standardize the process of CTG evaluation within the delivery ward.

Introduction

Cardiotocography (CTG), measurement of fetal heart rate and uterine contractions, remains the main instrument for direct assessment of the well-being of a fetus during delivery. The rationale of CTG monitoring is that it enables clinicians to intervene in a timely fashion in order to prevent potential adverse outcome of the labour. Nevertheless, since the very introduction of CTG in the 1960s, its merit was widely disputed [1]. The interpretation of CTG still suffers from large inter-observer disagreement, cf. [2–5], despite the introduction of clinical guidelines for CTG evaluation [e.g. the most prominent worldwide being the International Federation of Gynecology and Obstetrics (FIGO)] [6]. High inter- and intra-observer variability is reported frequently [5] (see also references therein for overview). According to a recent article by de Campos and Bernardes [7], guidelines are in general too complex with too many parameters, which are difficult to assess precisely in the clinical setting. Changes to the FIGO guidelines had been proposed extensively but still no general agreement on the guidelines exists. More complex guidelines were proposed by Parer *et al*. [8,9] and compared in terms of sensitivity and specificity to neonatal status at birth (measured by umbilical artery pH) with sometimes controversial results [10–12].

Since the CTG introduction, a large attention was devoted to analysis of inter- and intra-observer variability of its interpretation. The most recent study [5] shows only moderate agreement on CTG evaluation among obstetricians and is consistent with studies published before. In general, two possible solutions to overcome difficulties with CTG interpretation were suggested: more education and training [13] or cost-effective solution in form of decision support system serving as an aid and providing additional information [14–17]. Majority of these solutions still waits to be adopted into daily clinical practice. The only systems being used (mostly in 'research' conditions) are based on the FIGO parameters emulating and objectifying the decision-making process of clinicians. Among the most widespread systems are the Omniview-Sisporto [18], K2 System [19], Peri-CALM [20] and Monako [21].

This study provides contribution through an in-depth analysis of the CTG evaluation performed by nine experienced obstetricians from six obstetrics and gynecology clinics in the Czech Republic.

The analysis is based on following measures: proportion of classes assigned (normal, suspicious and pathological), time needed to decide on a class and intra-observer agreement. Results of clinicians' prediction based on the intrapartum CTG recording to umbilical artery pH are presented for the first time. Additionally, the study compares clinical evaluation (acquired using the CTGAnnotator software) and evaluation of attending obstetrician (acquired during delivery) in terms of proportion of agreement (PA), sensitivity and specificity. Finally, all the evaluations are available online together with the CTU-UHB database enabling further analysis.

Materials and methods

CTG database

Open access intrapartum CTG database [22] (the CTU-UHB database) is used as the set for the annotation. The database contains 552 CTG recordings selected from hospital repository of 9164 recordings based on clinical and technical parameters. Clinical parameters were used to achieve consistent database. Only singleton pregnancies and fetuses with more than 37 completed weeks of gestation and were included. Excluded were all fetuses with known intrauterine growth restriction, fetal infection and fetuses with congenital malformations. In addition, only records with second stage shorter than 30 minutes and distance of delivery to the end of CTG trace shorter than 30 minutes were selected for the database. In all recordings, two 30-minute windows of CTG in the first stage of labour and one window in the second stage of labour (if the second stage was available) were prepared. The full description of database is provided in Chudáček *et al*. [22]. The characteristics [mean (standard deviation)] of recordings are as follows: distance of the time of the end of the actual CTG signal and time of birth of 3 (3) in minutes; length of the first stage of labour of 225 (111) minutes; length of the second stage of labour 12 (8) minutes. CTGs with umbilical artery $\text{pH} \le 7.05$ were considered as pathological (38 cases), with $\text{pH} > 7.05$ and $\text{pH} \le 7.15$ as suspicious (56 cases) and the rest with $\text{pH} > 7.15$ as normal (412 cases).

Ethical committee approval to collect and work on the anonymized data had been granted and informed consents were obtained.

Data preparation for intra-observer agreement analysis

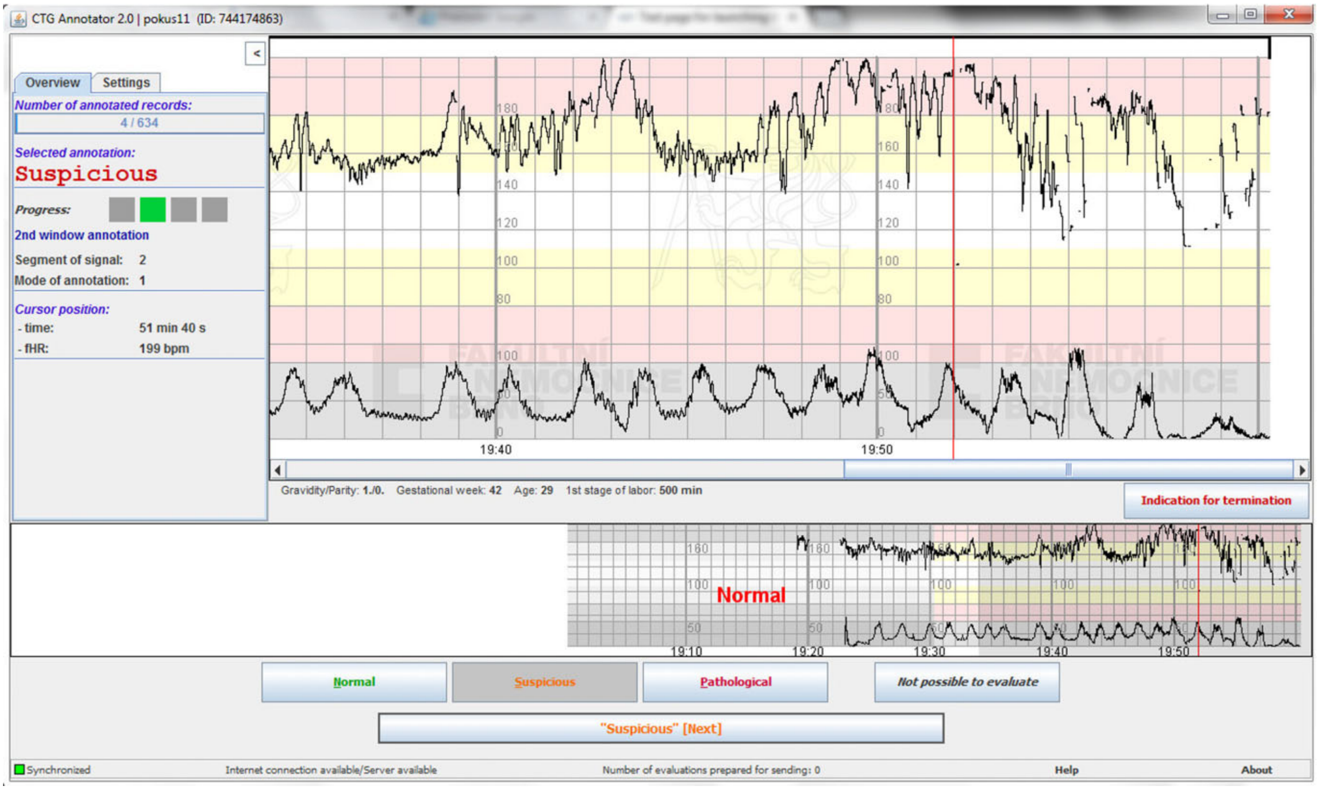
In order to establish intra-observer agreement, 82 recordings were randomly selected from the CTU-UHB database and were presented repeatedly to each clinician. The first half of repeated recordings (41 out of the 82 recordings) was selected randomly in advance and was kept the same for each user. The second half (another 41 recordings) was based on evaluation of each obstetrician and automatically selected by the application to acquire the same number of repeated evaluation for each evaluation class (normal, suspicious, pathological).

Annotation software

The CTGAnnotator [23], see screenshot in Fig. 1, was used to obtain annotation of the CTG recordings from nine obstetricians. The CTGAnnotator is a stand-alone Java-based application that was deployed separately at each clinic involved. The application adopted the most commonly used display layout of CTG machines (in European format – 1 cm min^{-1} and 20 bpm cm^{-1}) and therefore posed no difficulty for obstetricians to adjust. It presented the CTG trace in the form of consecutive 30-minute windows together with basic clinical information (described in Evaluation methodology section). The application was designed so that CTG recordings presented repeatedly to obstetricians were as far apart as possible from a first occurrence.

Clinical annotation

All obstetricians working on delivery wards of six obstetrics and gynecology departments of all the University Hospitals in the Czech Republic have been currently practising delivery ward doctors with median experience of 15 years (minimum 10, maximum 33). Simple introduction to the application was provided individually to each obstetrician at their workplace. It included a tutorial mode using additional CTG recordings that were not used for evaluation. It is expected that all obstetricians adhere to the FIGO guidelines criteria (as required for the clinical decision making by the official Czech Society of Gynecology and Obstetrics), any training was neither provided nor recommended.



CTG Annotator 2.0 | pokus11 (ID: 744174863)

Overview | Settings

Number of annotated records: 4 / 634

Selected annotation: Suspicious

Progress:

2nd window annotation

Segment of signal: 2

Mode of annotation: 1

Cursor position:

- time: 51 min 40 s

- fHR: 199 bpm

Gravidity/Parity: 1./0. Gestational week: 42 Age: 29 1st stage of labor: 500 min

Indication for termination

Normal | Suspicious | Pathological | Not possible to evaluate

"Suspicious" [Next]

Synchronized

Internet connection available/Server available

Number of evaluations prepared for sending: 0

Help

About

Figure 1 Cardiotocogram (CTG) recording and an example screen of the CTGAnnotator software that was used for CTG annotation.

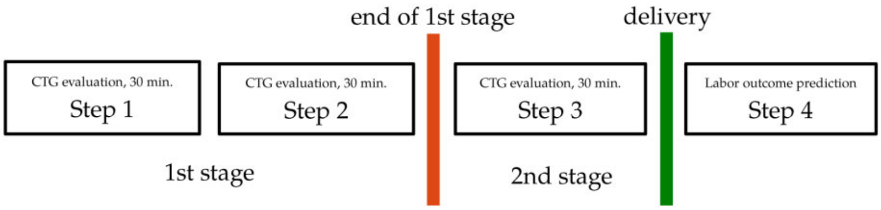


Figure 2 Workflow of the annotation process. Two 30-minute windows were evaluated in the first stage of labour and one window in the second stage (if the CTG signal was longer than 5 minutes). The labour outcome was evaluated in the step 4.

Based on the data structure each recording was presented in four steps, see Fig. 2:

1 30 minutes long window beginning at maximum 1 hour before the end of the first stage of labour

2 30 minutes long window beginning at maximum 30 minutes before the end of the first stage of labour

3 Full second stage of labour signal presented for evaluation if 5 minutes or more of CTG signal was available

4 Evaluation of labour outcome – prediction of umbilical artery pH after delivery

General clinical information about mothers' age, gestational age, gravidity, parity and total length of the first stage was provided. The obstetricians were aware *a priori* of the general outlines of the database as described in the previous section and the way of the presentation of the recordings.

From step 1 to 3, the clinicians evaluated CTG recordings as normal, suspicious, pathological or uninterpretable according to their daily practice. In step 4, clinicians predicted a labour outcome (umbilical artery pH after delivery), divided into four categories as no hypoxia (normal), mild hypoxia (suspicious), severe hypoxia (pathological) or uninterpretable. During introduction of CTGAnnotator these classes were described in terms of pH values (pH > 7.15 for no hypoxia; 7.15 ≥ pH > 7.05 for mild hypoxia and pH ≤ 7.05 for severe hypoxia). For the step 4, occurrences of risk factors, as written in the patient record, were provided in addition to the basic clinical information described earlier. It included presence of diabetes, fever, hypertension, preeclampsia, meconium stained amniotic fluid or induction of labour.

Evaluation methodology

The analysis of proportion of evaluation to each class, intra-/inter-observer agreement and sensitivity and specificity to objective outcome [pH and base deficit (BD)] is performed for the all steps of evaluation. If not stated otherwise, the results are presented as summary of all clinicians, for example for each clinicians, $i = 1, 2, \ldots, 9$, the proportion of evaluation to normal, suspicious and pathological class, n_i, s_i and p_i is computed, respectively. The

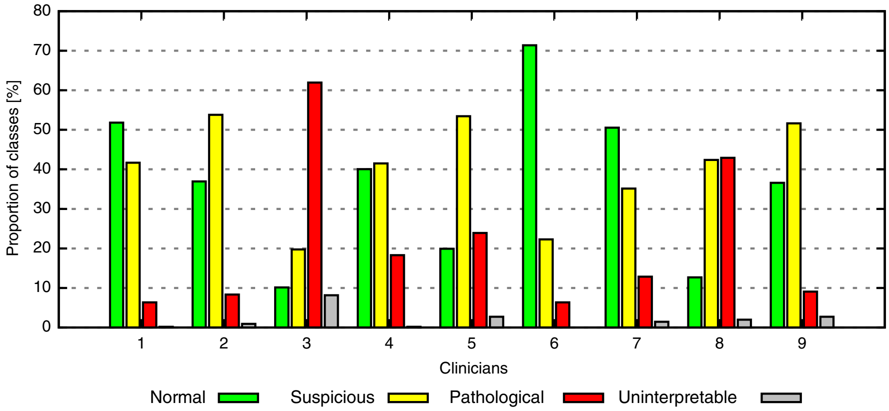


Figure 3 Proportion of evaluation by each obstetrician (1–9) into normal, suspicious, pathological and uninterpretable class based on 552 recordings evaluated.

summary results are presented as *median* (n_i), *median* (s_i) etc. Furthermore, the step 2 is used for detailed analysis for two reasons. First, step 1 is well reflected in the data of step 2. Second, step 3 (evaluation of second stage) is ridden with issues of signal quality because of increased maternal activity (93 CTG recordings has more than 50% of missing data) and is usually evaluated differently from the recording in the first stage of labour [24].

Proportions of recordings assigned by each clinician to normal, suspicious, pathological and non-interpretable CTG categories in step 2 were examined. We have also measured time in seconds needed to arrive at decision for each recordings (in the CTGAnnotator). For the step 4, predictions of the newborn outcome based on the CTG recording in step 1 to step 3 were evaluated with respect to the umbilical artery pH (pH) and BD computed according to Siggaard-Andersen and Huch [25]. For 296 CTG recordings,[1] the CTG evaluations (normal, suspicious and pathological) were available directly from the paper at Masaryk University Hospital Brno. These recordings were evaluated by an attending doctor on a printed CTG trace (as part of the patient documentation) during delivery. Such data represent unique insight into the evaluation of the CTG in daily clinical practice. Results based on these annotations are analysed with respect to the pH and BD in order to obtain sensitivity and specificity.

Statistical analysis

The results are presented as a median with 95% confidence intervals (CIs) estimated using bias-corrected and accelerated bootstrap method [26]. The Fleiss kappa coefficient [27] and proportion of agreement (PA) are used for analysis of agreement. The kappa coefficient is predominantly used to measure agreement in other studies, cf. [2,3,5,28], but it is heavily dependent on observed marginal proportions (making comparison across different population hardly possible) and also lacks natural extension for multiple rates and multinomial classes [29,30]. Thus, we consider the kappa coefficient as only a complementary measure. The Kruskal–Wallis test is used for statistical evaluation; $P < 0.05$ indicates statistical significance.

Table 1 Median proportions of evaluation to each class

Evaluation	Step 1	Step 2	Step 3	Step 4
Normal	49	37	29	68
Suspicious	41	42	37	24
Pathological	7	13	17	7
Uninterpretable	1	1	4	2

Results

Nine clinicians evaluated 552 CTG recordings. From these recordings, 82 were repeated (for intra-observer agreement) making it 634 CTGs evaluated in total by each clinician. Clinicians were randomly assigned a number, which can be used to connect the respective results across all the figures.

Analysis of individual clinicians (proportion of evaluation, intra-observer agreement)

Table 1 shows proportion of evaluation into different classes and confirms that clinicians evaluate CTG dominantly as normal and suspicious. Interestingly, from step 1 to step 3 (i.e. longer CTG recording for decision), the proportion of normal class is redistributed in favour to suspicious and pathological class. The detailed analysis for step 2, cf. Fig. 3, clearly shows that proportion of normal, suspicious, pathological and uninterpretable annotation differs substantially for each clinician, clinicians 3 and 8 are the most defensive ones (higher number of pathological evaluation), clinician 6, on the other hand, is the most confident one.

Table 2 shows median time in seconds of evaluation for each step and class, with pathological/uninterpretable class evaluation taking the longest and normal evaluation the shortest time overall. Details for step 2 and individual clinicians are shown in Fig. 4. All clinicians had significantly different time for each type of evaluation on significance level $P < 0.05$ (Kruskal–Wallis). The median length of annotation, that is time elapsed between evaluation of the first and last recording, was 38 days. The repeated CTG recordings (for intra-observer agreement) were presented on average 5 days after the first occurrence. Hence, it is unlikely that clinicians would remember the repeated CTGs.

The median of intra-observer PA for all clinicians was about 70% for the all steps. Table 3 shows slightly better agreement for

[1] Only those recordings with CTG evaluation for the steps (1–3) were selected. This selection was performed prior to the experiment described in this paper.

step 1 (76%) although it could be linked to higher proportion of normal class in step 1, cf. Table 1 for details. The detailed results for step 2 are presented in Fig. 5. Two clinicians (3 and 8) have large agreement on pathological evaluation in comparison with other clinicians. This can be explained by disproportionate amount of pathological evaluations, as it is evident in Fig. 2.

Table 2 Median time for evaluation into different classes

Evaluation	Step 1	Step 2	Step 3	Step 4
Normal	3.4	2.9	3.4	3.9
Suspicious	5.8	4.5	4.6	4.3
Pathological	6.4	6.3	5.1	8
Uninterpretable	7.5	5.7	4.5	7.3
All	4.4	3.6	4.1	4.5

Table 3 Median intra-observer proportion of agreement

Evaluation	Step 1	Step 2	Step 3	Step 4
Normal	86	77	69	85
Suspicious	69	61	61	59
Pathological	57	59	63	52
Uninterpretable	0	0	40	0
All	76	70	71	71

Analysis of inter-observer agreement

The inter-observer agreement, cf. Table 4, follows similar patterns across all steps as the intra-observer agreement. The agreement on normal class decreases from step 1 to step 3, which could be linked to the decrease in proportion of normal evaluation in Table 1.

The reason for the low PA value is that two clinicians (3 and 8) could be considered as outliers as they evaluated CTG defensively (higher proportion of pathological), as it was shown in Fig. 2. This is also clearly presented in Fig. 6, where the pairwise PA is shown in form of symmetric matrix. The majority voting outcome based on evaluation of all clinicians is shown in the first row and first column labelled by 0. The rest of the rows and columns represent individual clinicians as in the previous figures. For example, the first clinician (marked with 1) agrees with majority voting (marked with 0) on 80% and with second clinician (marked with 2) on 65%. Again, it can be seen that clinicians 3 and 8 are distinct from the others. The overall kappa was: 0.255 with 95% CI (0.253–0.258) being the fair agreement although we warn against its improper use to compare different populations.

Clinical evaluation to pH and to the CTG evaluation stored in hospital records

Table 5 presents sensitivity and specificity of the majority vote of the nine clinicians with respect to the objective outcome measures

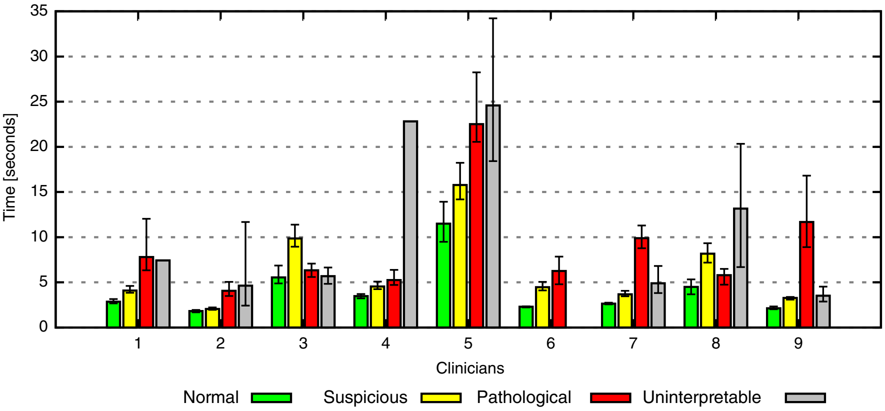


Figure 4 Median time for evaluation into different classes with 95% confidence intervals.

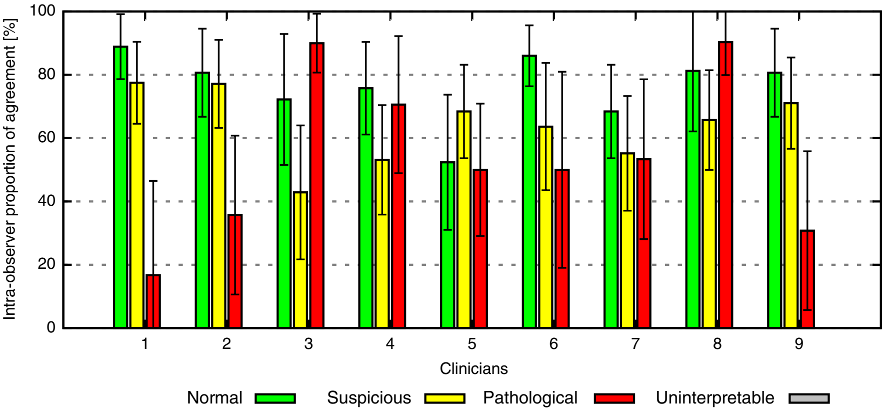


Figure 5 Intra-observer proportion of agreement based on the analysis of the repeated recordings (82 in total) with 95% confidence intervals in respect to assessed CTG class.

(pH, BD and Apgar score at fifth minute). To be able to compute sensitivity and specificity, and for the sake of clarity, the results are presented as evaluation of pathological versus suspicious + normal traces. From step 1 to step 3, the sensitivity increases while the specificity decreases for all measures. For the step 4 (labour outcome prediction), the sensitivity decreases again. The distribution of pH with respect to labour evaluation is shown in Fig. 7. Ideally, for recordings with pH ≤ 7.05, there should be only evaluation to pathological class. However, the pathological evaluation is scattered across the whole range of pH.

Table 4 Inter-observer proportion of agreement with 95% confidence interval

Evaluation	Step 1	Step 2	Step 3	Step 4
Normal	69 (66–72)	57 (54–60)	48 (44–52)	65 (63–67)
Suspicious	49 (47–51)	46 (44–48)	41 (39–43)	32 (30–34)
Pathological	33 (28–39)	41 (36–46)	42 (37–47)	29 (24–33)
Uninterpretable	26 (14–44)	15 (10–20	26 (20–33)	20 (16–24)
All	56 (54–58)	48 (47–50)	63 (60–66)	50 (48–52)

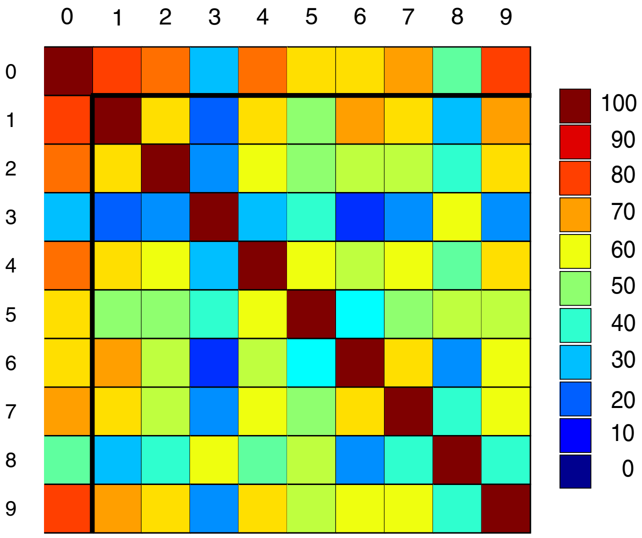


Figure 6 Matrix of inter-observer agreement for step 2. Majority voting of all clinicians is marked by 0. Clinicians are marked by number 1–9. The matrix represents pairwise comparison of proportion of agreement between all clinicians.

CTG evaluation from hospital records

The PA between evaluation of the first stage from the clinical documentation and the evaluation using the majority voting of nine obstetricians was 59%. The sensitivity and specificity for majority voting and hospital records, cf. Table 6, are remarkably similar to each other.

Discussion and conclusion

This study presents the analysis of CTG evaluation by nine experienced obstetricians. Each clinician evaluated 634 recordings (approximately 691 hours of CTG recordings). It is the largest study ever performed when both the number of clinicians and number of recordings are considered, providing broader insight into clinical evaluation and its variability. In contrast to other studies published in medical journals, it is not restricted to a quantitative measure of agreement such as kappa coefficient but it aims to provide a clear picture of clinicians' agreement/disagreement. The analysis is performed on well-constructed and documented open access database (CTU-UHB database). The clinical evaluations are available upon request and are published online at webpage of corresponding author.

The assessment of agreement between observers is not an easy task. Among statisticians, there is no accordance on how the observer agreement should be measured. For the categorical variables, the kappa coefficient has been used in the past to measure agreement in many studies, for example [2,3,5,28]. But it has been shown that the kappa is dependent on observed marginal proportions making comparison across different population infeasible. Additionally, the kappa coefficient lacks natural extension for multiple rates and multinomial classes [29,30]. In general, there is no single measure of agreement that could outperform the others [31]; therefore, the PA and also Fleiss kappa value [27] were used. However, having in mind disadvantages of the kappa coefficient, it is considered as a complementary measure.

The analysis of CTG evaluation of individual clinicians revealed that each clinician has different proportion of assigned classes. Clinicians mostly evaluated CTG as normal and suspicious. Time needed to decide on CTG evaluation was approximately the same for all clinicians. Clinicians spent more time with decision on pathological class than on normal and suspicious. This was expected as pathological recordings could be more complex and difficult to evaluate.

The presented large intra- and inter-observer variability supports the results of previous studies [4,5]. Obstetricians using FIGO

Table 5 Sensitivity and specificity for majority voting of nine clinicians. The results are presented for the 552 records (pathological vs. normal + suspicious evaluation)

	Step 1		Step 2		Step 3		Step 4	
Objective	SE	SP	SE	SP	SE	SP	SE	SP
pH ≤ 7.05	12 (5–26)	95 (92–96)	27 (16–42)	89 (86–92)	59 (38–78)	87 (82–90)	25 (13–41)	95 (93–97)
BD ≥ 12	18 (5–42)	94 (92–96)	39 (16–63)	89 (86–91)	43 (13–77)	84 (80–88)	24 (8–49)	95 (92–96)
Apgar < 7	11 (2–33)	94 (92–96)	11 (2–32)	88 (85–91)	39 (17–66)	85 (80–88)	21 (8–45)	95 (92–96)

BD, base deficit; SE, sensitivity; SP, specificity.

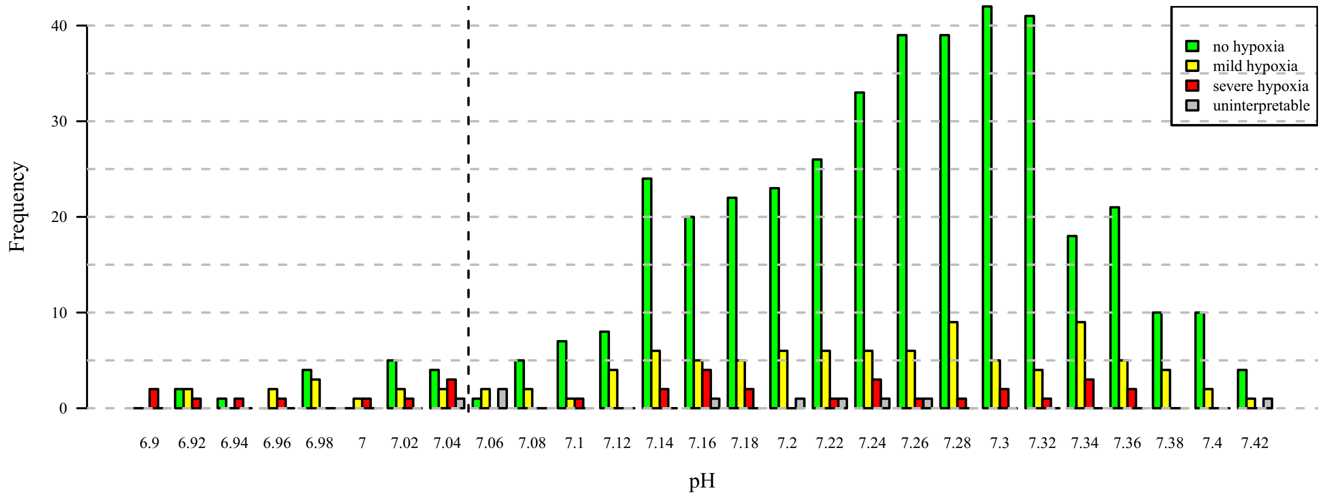


Figure 7 Majority voting versus pH (step 4). The vertical line marks pH threshold (pH ≤ 7.05).

Table 6 Sensitivity and specificity of CTG evaluation as noted in hospital records and for majority voting based on nine clinicians. The results are presented for the same 296 CTG recordings (pathological vs. normal + suspicious evaluation)

		Step 1		Step 2		Step 3		Step 4	
Annotation	Objective	SE	SP	SE	SP	SE	SP	SE	SP
Majority voting	pH ≤ 7.05	29 (12–54)	92 (88–95)	41 (20–65)	86 (81–90)	86 (45–99)	86 (79–90)	38 (18–63)	94 (91–97)
	BD ≥ 12	30 (9–62)	92 (87–95)	50 (22–78)	86 (81–90)	50 (3–94)	83 (76–88)	22 (4–56)	93 (89–96)
	Apgar < 7	50 (10–90)	91 (87–94)	50 (10–90)	85 (80–89)	100 (5–100)	83 (76–89)	75 (25–99)	93 (90–96)
Hospital records	pH ≤ 7.05	41 (20–65)	94 (91–97)	41 (20–65)	94 (90–97)	40 (15–71)	93 (88–96)	N/A	N/A
	BD ≥ 12	60 (29–85)	94 (91–97)	60 (29–85)	94 (90–96)	25 (1–75)	92 (87–95)	N/A	N/A
	Apgar < 7	0 (0–53)	92 (88–95)	0 (0–53)	92 (87–95)	33 (2–86)	92 (87–95)	N/A	N/A

BD, base deficit; CTG; N/A, Not Available; SE, sensitivity; SP, specificity.

guidelines or its derivatives struggle with consistency of their assessment of CTG recordings regardless of the experience or academic stature. The intra-observer agreement of 71% for step 2 was partly caused by lower agreement on pathological class (59%). For the normal class, the intra-observer agreement was 77%.

The PA in step 2 was 48%. There are several reasons for low PA. Two clinicians (3 and 8) provided distinct evaluation and the PA computed without these clinicians raised significantly to 58%. Another possible factor is lower inter-observer agreement on pathological cases, cf. Table 4. Our recent work [32] suggests that suspicious and pathological classes are broad, possibly not well defined, and that the boundary between suspicious and pathological cases is different for each clinician. The direct comparison of inter- and intra-observer agreement to the state-of-the-art publications is rather difficult. First, the number of recordings evaluated in other studies ranges from 3 [33] to 845 [3] with 30 to 50 recordings being the most common size, for example [2,4,34]. Regarding the number of evaluators, again, the wide range and professional background of the experts can be found from 116 Obstetrics and gynaecology (ob-gyn) residents in Lidegaard *et al*. [35], 28 midwives in Devane and Lalor [33] to three experts in Ojala *et al*. [36]. Second, inter-observer agreement is usually presented using kappa coefficient, which cannot be compared across different populations.

In general, fetal heart rate has high sensitivity and low specificity to fetal acidosis. However, in our study, the sensitivity and specificity of clinicians' evaluation of labour outcome (step 4) to fetal acidosis was 25% (95% CI 13–41) and 95% (CI 93–97), respectively. The clinical evaluation has lower sensitivity than specificity. These results are consistent with those reported for the National Institute of Child Health and Human Development classification system in Coletta *et al*. and Tommaso *et al* [10,12]. For the FIGO guidelines, the sensitivity and specificity was evaluated only with a computerized system for CTG evaluation [37]. In Schiermeier *et al*. [37], the computerized system had higher sensitivity and lower specificity. We would like to note here that comparison of these studies [10,12,37] is limited as each study aimed at different scenario and chose a different pH threshold.

The reported sensitivity and specificity in Table 6 is remarkably similar to the sensitivity of 41% and specificity of 94% of the CTG evaluation recorded by an on-duty obstetrician on printed traces during the actual delivery. This implies that the process of annotation and information provided to clinicians corresponds to the situation in clinical practice.

This work provides expert obstetrician's annotations for the CTU-UHB database that are open to research community, available at Physionet [22]. This paper brings up several topics which are of broader interest. It is well known that interpretation of CTG is accompanied by high intra-/inter-observer variability as it has been reported from the beginnings of the CTG. The solutions to improve the interpretation were interestingly summarized and discussed in Santo and Ayres-de-Campos [38]. The very first step forward is to simplify guidelines for interpretation and keep them as simple and as objective as possible. The FIGO guidelines were not updated since its introduction in 1986 [6] and are partly in disagreement with several national guidelines, cf. [7] for guidelines comparison. An alternative to guidelines simplification (or its implementation in clinical practice) would be to employ intensive education (training and testing sessions) [19,39]. However, the regular education is very resource demanding (especially time) and thus usually sustained only for limited time periods after which the additional knowledge deteriorates. Another possibility is to use a computer-based interpretation of CTG although algorithmic implementation of FIGO guidelines is cumbersome as the guidelines offers room for different interpretation, for example the circular definition of baseline heart rate and acceleration and deceleration. It is not surprising that computer interpretation of FIGO criteria, implemented and reported in Schiermeier *et al.* [37], and its sensitivity and specificity to pH is in strong contradiction to the sensitivity and specificity achieved by clinicians in our study – the FIGO guidelines are simply not rigorously followed in clinical practice.

From the perspective of computer-based CTG interpretation, two directions are possible: to design a system adhering strictly to guidelines [37] and use only morphological CTG features defined in the guidelines or to take advantage of ongoing research in fetal heart rate variability [14,16,40–42] and use features that probe the FHR in more detail and thus go beyond FIGO. In this case, the clinical evaluation of CTG together with pH and BD could be used to guide the choice of new features. Either way, the best possible approach is to implement all the steps to improve the CTG evaluation, that is to make the guidelines as simple and as objective as possible, to establish and maintain education of clinical personal and to employ computer interpretation of CTG that would aid clinicians' decisions.

Acknowledgment

The presented work was partially funded by Ministry of Healthcare of the Czech Republic grant no. NT11124-6/2010 and SGS grant of the CTU SGS13/203/-OHK3/3T/13 and by Czech Grant Agency project number: 14-28462P Statistical methods of intrapartum CTG signal processing in the context of clinical information. Additionally, we would like to thank Lukáš Zach for his development work on the CTGAnnotator, Philips Healthcare for providing us with the tool to be able convert the data from the OB TraceVue system to readable format and local distributor representative Ms. Monika Jiráčková from S&T for providing us with her expertise.

References

1. Sartwelle, T. P. (2012) Electronic fetal monitoring: a bridge too far. *The Journal of Legal Medicine*, 33 (3), 313–379.
2. Bernardes, J., Costa-Pereira, A., de Campos, D. A., van Geijn, H. P. & Pereira-Leite, L. (1997) Evaluation of interobserver agreement of cardiotocograms. *International Journal of Gynaecology and Obstetrics*, 57 (1), 33–37.
3. Blix, E., Sviggum, O., Koss, K. S. & Oian, P. (2003) Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts. *BJOG: An International Journal of Obstetrics and Gynaecology*, 110 (1), 1–5.
4. Vayssiere, C., Tsatsaris, V., Pirrello, O., Cristini, C., Arnaud, C. & Goffinet, F. (2009) Inter-observer agreement in clinical decision-making for abnormal cardiotocogram (CTG) during labour: a comparison between CTG and CTG plus STAN. *BJOG: An International Journal of Obstetrics and Gynaecology*, 116 (8), 1081–1087, discussion 1087–8.
5. Blackwell, S. C., Grobman, W. A., Antoniewicz, L., Hutchinson, M. & Gyamfi Bannerman, C. (2011) Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System. *American Journal of Obstetrics and Gynecology*, 205 (4), 378.e1–378.e5.
6. FIGO (1986) Guidelines for the use of fetal monitoring. *International Journal of Gynecology & Obstetrics*, 25, 159–167.
7. de Campos, D. A. & Bernardes, J. (2010) Twenty-five years after the FIGO guidelines for the use of fetal monitoring: time for a simplified approach? *International Journal of Gynecology & Obstetrics*, 110 (1), 1–6.
8. Parer, J. T. & Ikeda, T. (2007) A framework for standardized management of intrapartum fetal heart rate patterns. *American Journal of Obstetrics and Gynecology*, 197 (1), 26.e1–26.e6.
9. Parer, J. T., Ikeda, T. & King, T. L. (2009) The 2008 National Institute of Child Health and Human Development report on fetal heart rate monitoring. *Obstetrics and Gynecology*, 114 (1), 136–138.
10. Coletta, J., Murphy, E., Rubeo, Z. & Gyamfi-Bannerman, C. (2012) The 5-tier system of assessing fetal heart rate tracings is superior to the 3-tier system in identifying fetal acidemia. *American Journal of Obstetrics and Gynecology*, 206 (3), 226.e1–226.e5.
11. Miller, D. A. & Miller, L. A. (2012) Three-tier versus five-tier fetal heart rate classification systems. *American Journal of Obstetrics and Gynecology*, 207 (6), e8–e9, author reply e9.
12. Tommaso, M. D., Seravalli, V., Cordisco, A., Consorti, G., Mecacci, F. & Rizzello, F. (2013) Comparison of five classification systems for interpreting electronic fetal monitoring in predicting neonatal status at birth. *The Journal of Maternal-fetal and Neonatal Medicine*, 26 (5), 487–490.
13. Doria, V., Papageorghiou, A. T., Gustafsson, A., Ugwumadu, A., Farrer, K. & Arulkumaran, S. (2007) Review of the first 1502 cases of ECG-ST waveform analysis during labour in a teaching hospital. *BJOG: An International Journal of Obstetrics and Gynaecology*, 114 (10), 1202–1207.
14. Bernardes, J. & De-Campos, D. A. (2010) The persistent challenge of foetal heart rate monitoring. *Current Opinion in Obstetrics and Gynecology*, 22 (2), 104–109.
15. Hasley, S. K. (2011) Decision support and patient safety: the time has come. *American Journal of Obstetrics and Gynecology*, 204 (6), 461–465.
16. Hamilton, E. F. & Warrick, P. A. (2013) New perspectives in electronic fetal surveillance. *Journal of Perinatal Medicine*, 41 (1), 83–92.
17. Jezewski, J., Wrobel, J., Horoba, K., Gacek, A. & Sikora, J. (2002) "Fetal heart rate variability: clinical experts versus computerized system interpretation." Engineering in Medicine and Biology, 2002. 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society EMBS/BMES Conference, 2002. Proceedings of the Second Joint, vol. 2, no., pp. 1617, 1618 vol. 2, 2002 doi: 10.1109/IEMBS.2002.1106566, pp. 1617–1618. Houston: IEEE.

18. de Campos, D. A., Sousa, P., Costa, A. & Bernardes, J. (2008) Omniview-SisPorto® 3.5 – a central fetal monitoring station with online alerts based on computerized cardiotocogram + ST event analysis. *Journal of Perinatal Medicine*, 36 (3), 260–264.
19. Beckley, S., Stenhouse, E. & Greene, K. (2000) The development and evaluation of a computer-assisted teaching programme for intrapartum fetal monitoring. *BJOG: An International Journal of Obstetrics and Gynaecology*, 107 (9), 1138–1144.
20. Elliott, C., Warrick, P. A., Graham, E. & Hamilton, E. F. (2010) Graded classification of fetal heart rate tracings: association with neonatal metabolic acidosis and neurologic morbidity. *American Journal of Obstetrics and Gynecology*, 202 (3), 258.e1–258.e8.
21. Jezewski, J., Wrobel, J., Horobal, K., Kupka, T. & Matonia, A. (2006) "Centralised fetal monitoring system with hardware-based data flow control," Advances in Medical, Signal and Information Processing, MEDSIP 2006 IET 3rd International Conference On, vol., no., Publisher: IEEE, pp. 1–4. 17–19 July 2006.
22. Chudáček, V., Spilka, J., Burša, M., *et al*. (2014) Open access intrapartum CTG database. *BMC Pregnancy and Childbirth*, 14 (1), 16.
23. Zach, L., Chudáček, V., Huptych, M., Spilka, J., Burša, M. & Lhotská, L. (2013) CTG Annotator-novel tool for better insight into expert-obstetrician decision making processes. In World Congress on Medical Physics and Biomedical Engineering May 26–31, 2012 (ed. Long, M.), pp. 1280–1282. Beijing, China: Springer Berlin Heidelberg.
24. Tranquilli, A. L. (2012) Fetal heart rate in the second stage of labor: recording, reading, interpreting and acting. *Journal of Maternal-Fetal & Neonatal Medicine*, 25 (12), 2551–2554.
25. Siggaard-Andersen, O. & Huch, R. (1995) The oxygen status of fetal blood. *Acta Anaesthesiologica Scandinavica Supplementum*, 107, 129–135.
26. Efron, B. (1994) Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89 (426), 463–475.
27. Fleiss, J. L., Levin, B. & Paik, M. C. (2004) Statistical Methods for Rates and Proportions. Hoboken, NJ: John Wiley & Sons, Inc.
28. Costa-Santos, C., Bernardes, J., Antunes, L. & Ayres-de-Campos, D. (2011) Complexity and categorical analysis may improve the interpretation of agreement studies using continuous variables. *Journal of Evaluation in Clinical Practice*, 17 (3), 511–514.
29. Feinstein, A. R. & Cicchetti, D. V. (1990) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43 (6), 543–549.
30. Cicchetti, D. V. & Feinstein, A. R. (1990) High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43 (6), 551–558.
31. Santos, C. C., Pereira, A. C. & Bernardes, J. (2005) Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences. *BJOG: An International Journal of Obstetrics and Gynaecology*, 112 (5), 667–669.
32. Spilka, J., Chudáček, V., Janků, P., *et al*. (2014) Analysis of obstetricians' decision making on CTG recordings. *Journal of Biomedical Informatics*, 51 (0), 72–79.
33. Devane, D. & Lalor, J. (2005) Midwives' visual interpretation of intrapartum cardiotocographs: intra- and inter-observer agreement. *Journal of Advanced Nursing*, 52 (2), 133–141.
34. Keith, R. D., Beckley, S., Garibaldi, J. M., Westgate, J. A., Ifeachor, E. C. & Greene, K. R. (1995) A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *British Journal of Obstetrics and Gynaecology*, 102 (9), 688–700.
35. Lidegaard, O., Bøttcher, L. M. & Weber, T. (1992) Description, evaluation and clinical decision making according to various fetal heart rate patterns. Inter-observer and regional variability. *Acta Obstetricia et Gynecologica Scandinavica*, 71 (1), 48–53.
36. Ojala, K., Mäkikallio, K., Haapsamo, M., Ijäs, H. & Tekay, A. (2008) Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies. *Acta Obstetricia et Gynecologica Scandinavica*, 87 (5), 536–540.
37. Schiermeier, S., von Steinburg, S. P., Thieme, A., *et al*. (2008) Sensitivity and specificity of intrapartum computerised FIGO criteria for cardiotocography and fetal scalp pH during labour: multicentre, observational study. *BJOG: An International Journal of Obstetrics and Gynaecology*, 115 (12), 1557–1563.
38. Santo, S. & Ayres-de-Campos, D. (2012) Human factors affecting the interpretation of fetal heart rate tracings: an update. *Current Opinion in Obstetrics and Gynecology*, 24 (2), 84–88.
39. Pehrson, C., Sorensen, J. L. & Amer-Wåhlin, I. (2011) Evaluation and impact of cardiotocography training programmes: a systematic review. *BJOG: An International Journal of Obstetrics and Gynaecology*, 118 (8), 926–935.
40. Henriques, T., Gonçalves, H., Antunes, L., Matias, M., Bernardes, J. & Costa-Santos, C. (2013) Entropy and compression: two measures of complexity. *Journal of Evaluation in Clinical Practice*, 19 (6), 1101–1106.
41. Chudáček, V., Anden, J., Mallat, S., Abry, P. & Doret, M. (2014) Scattering transform for intrapartum fetal heart rate variability fractal analysis: a case-control study. *IEEE Transactions Biomedical Engineering*, 61 (4), 1100–1108.
42. Georgieva, A., Papageorghiou, A. T., Payne, S. J., Moulden, M. & Redman, C. W. G. (2014) Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth. *BJOG: An International Journal of Obstetrics and Gynaecology*, 121 (7), 889–894.