Analysis of obstetricians' decision making on CTG recordings

Jiří Spilka[a,*], Václav Chudáček[a], Petr Janků[b], Lukáš Hruban[b], Miroslav Burša[a], Michal Huptych[a], Lukáš Zach[a], Lenka Lhotská[a]

[a]*Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic*
[b]*Department of Gynecology and Obstetrics, Teaching Hospital of Masaryk University in Brno, Czech Republic*

Abstract

Interpretation of cardiotocogram (CTG) is a difficult task since its evaluation is complicated by a great inter- and intra-individual variability. Previous studies have predominantly analyzed clinicians' agreement on CTG evaluation based on quantitative measures (e.g. kappa coefficient) that do not offer any insight into clinical decision making. In this paper we aim to examine the agreement on evaluation in detail and provide data-driven analysis of clinical evaluation.

For this study, nine obstetricians provided clinical evaluation of 634 CTG recordings (each ca. 60 min long). We studied the agreement on evaluation and its dependence on the increasing number of clinicians involved in the final decision. We showed that despite of large number of clinicians the agreement on CTG evaluations is difficult to reach. The main reason is inherent inter- and intra-observer variability of CTG evaluation.

Latent class model provides better and more natural way to aggregate the CTG evaluation than the majority voting especially for larger number of clinicians. Significant improvement was reached in particular for the pathological evaluation – giving a new insight into the process of CTG evaluation. Further, the analysis of latent class model revealed that clinicians unconsciously use four classes when evaluating CTG recordings, despite the fact that the clinical evaluation was based on FIGO guidelines where three classes are defined.

Keywords: cardiotocography, fetal heart rate, observer variation, biomedical informatics, decision making, latent class analysis

1. Introduction

Interpretation of cardiotocogram (CTG). The CTG is a simultaneous recording of fetal heart rate (FHR) and uterine contractions. It is an integral part of every day clinical practice. However, since its introduction, it has been a subject of many controversies as well as malpractice litigations [1]. The evaluation of CTG is accompanied with high intra- and inter-observer variability from the very beginning. And even though guidelines, e.g. the most prominent FIGO guidelines [2], were introduced to tackle the heterogeneity of the CTG evaluation, high inter- and intra-observer variability is reported frequently even today [3].

The FIGO guidelines consists of 3-tier classification system and in 1980s became the first internationally recognized guidelines. Since then national alternatives with minor tweaks were introduced [4, 5, 6]. The comparison of various guidelines and their statements was performed by de Campos and Bernardes [7] with conclusion that the guidelines are, in general, too complex and hard to follow and thus attribute to high inter- and intra-observer variability. To better interpret the CTG patterns and to lower the variability additional improvements were suggested. Schifrin stated [8] that the guidelines lack a definition that can identify the transition from normal to ominous CTG – the so called conversion pattern. Parer and Ikeda [9] and Parer et al. [10] proposed an extension of the guidelines to a 5-tier system. A comparison in [11] claimed this system to be superior to the classical guidelines. Recently, Tommaso et al. showed [12] that the NICHD[1] guidelines have better sensitivity and specificity over 5-tier system. But in general, the performance of 5-tier was better since NICHD evaluated a lot of recordings as "intermediate". Further, Coletta et al. claimed [13] that there is better sensitivity using the 5-tier system, though the contrary was claimed in [14]. Despite all the efforts, none of the major guidelines changes were thoroughly evaluated in a larger group settings exceeding several hospitals interested.

Agreement on interpretation. The substantial inter- intra-observer variability makes it difficult to reach agreement on CTG interpretation. For the purpose of this paper, the agreement does not mean a discussion and consensus of all the clinicians in a consulting room. It means reaching an agreement over independently evaluated CTGs. Generally, the majority voting is a natural way to aggregate different opinions. When making decisions, people usually use weighted majority voting where weights are based on experience, reputation, work place, and other factors. However the determination of weights is subjective and could be misleading.

Observer agreement measures. Among statisticians there is no general agreement on how the observer agreement should be

*Corresponding author, Karlovo namesti 13, Prague 2, 121 35 Czech Republic, Phone: +420 224 357 325

Email address: spilka.jiri@fel.cvut.cz (Jiří Spilka)

[1] Eunice Kennedy Shriver National Institute of Child Health and Human Development

measured. The kappa coefficient, proportion of agreement, and intraclass correlation coefficient are the most used measures for agreement [15] even though they have many flaws. For example, the kappa coefficient is influenced by prevalence and base rate and is not suitable for comparison across different studies [16, 17]. Also it lacks a natural extension to multiple rates and multinomial classes. There is no single measure of agreement that could outperform the others [15]. The use of quantitative measures and reporting a single value of agreement is tempting, however the results are usually difficult to interpret.

Goals and contributions. In our work we aim at examining the agreement of obstetricians using latent class analysis and majority voting. In Section 2.1 we briefly describe the process of annotation that was performed on the CTU-UHB[2] database [19]. In Sections 2.3.1 and 2.3.2 we further describe the most common method to aggregate different opinions – the majority voting together with an alternative – the latent class analysis. In Section 3.1 we examine stability of clinicians' agreement using these two methods and show that the agreement is greatly improved especially on pathological class when using the latent class analysis. The latent class analysis shows us a different perspective on the controversial question of how many classes should be used for CTG evaluation. According to our results, the four class model yielded the best results, despite the fact, that clinicians had used guidelines with three classes (cf. Section 3.2).

2. Materials and methods

2.1. Clinical annotations

Evaluation of CTG recordings has been acquired using standalone application (*CTGAnnotator* [18]). The CTGAnnotator adopts the most commonly used display layout of CTG machines (in European format – 1 min/cm and 30 bpm/cm), and therefore poses no difficulty for clinicians to adjust. The evaluations were obtained from nine clinicians working on delivery wards of six Obstetrics and Gynaecology Clinics in the Czech Republic. All the clinicians are currently working in delivery practice with experience ranging from 10 to 33 years (with a median value of 15 years). The CTU-UHB intrapartum CTG database [19] was used for evaluation. All the experts had to undergo a basic training on the experiment methodology and the CTGAnnotator usage. Although we expected that all experts adhered to the FIGO guidelines criteria (as required by the Czech authorities[3]) we did not provide any special training for it nor we encouraged it. In our retrospective study we used evaluation of 60 min of CTG recordings at the end of the first stage of labor. Clinicians evaluated the CTG recordings into three classes: normal, suspicious, and pathological (FIGO classes).

2.2. Observer agreement

We use proportion of agreement (PA) to measure the agreement between clinicians. The PA is simply probability that clinicians agree on evaluation. We decided to use PA, which is intuitive and understandable, instead of other complex statistical measures that could obscure the analysis.

2.3. Voting schemes

The different schemes of voting were thoroughly studied in social sciences. The famous Condorcet's jury theorem (1786), details e.g. [20], states: if voters are right with probability $p > 1/2$, then majority vote is more likely to be right than wrong and the probability of being right tends to 1 when number of voters goes to infinity. Intuitively it is expected that potential variability could be cancelled out by a high number of voters.

Let y_i^j be a evaluation of the i-th example, $i = \{1, 2, \ldots, N\}$, given by the j-th clinician, $j = \{1, 2, \ldots, J\}$. Further let $c \in C$ be a category to which y_i^j could be assigned and $\delta(y_i^j, c)$ be an indicator function that equals 1 when the j-th clinician evaluates $y_i^j = c$ and 0 otherwise.

2.3.1. Majority voting

The majority voting is a simple voting mechanism to aggregate evaluation from J clinicians. The probability that the i-th example is assigned to the c-th class is

$$\mu_{ic} = \frac{1}{J} \sum_{j=1}^{J} \delta(y_i^j, c), \tag{1}$$

The majority voting, or more precisely plurality voting, is simply choosing a class c for maximum of μ_{ic}. In the case of ties a flip of fair coin is performed.

Problems with majority voting. Majority voting is simple and usually preferred method. However, there are some limitations when using majority voting on evaluation of CTG, which are summarized as follows:

1. There is high inter- and intra-observer variability in clinical evaluation (see for example [3, 22, 23]) and agreement might not be reached.
2. Each clinician has different expertise not only based on the length of his/her career (experienced vs. inexperienced) but also influenced by labor management at workplace; e.g. a clinician who is called only to the most serious cases could loose, to some extent, knowledge related to the normal cases.
3. Clinicians could loose concentration/motivation or be simply distracted during annotation.

2.3.2. Latent class analysis

The latent class analysis (LCA) is used to estimate the true (unknown/hidden) evaluation of CTG and to infer weights of individual clinicians' evaluation – the latent class model (LCM). Let $y_i \in \mathcal{Y}; \mathcal{Y} = \{1, 2, \ldots, C\}$ be the unobservable ground truth for the i-th example and $\boldsymbol{\alpha}_c^j = (\alpha_{c1}^j, \alpha_{c2}^j, \ldots, \alpha_{ck}^j, \ldots, \alpha_{cC}^j)$ be a multinomial parameter that represents probabilities that the c-th class corresponds to an evaluation in the k-th class, $k \in C$, assigned by the j-th clinician

$$\alpha_{ck}^j = P(y_i^j = k | y_i = c), \quad \alpha_{ck} \geq 0, \quad \sum_{k=1}^{C} \alpha_{ck}^j = 1. \tag{2}$$

[2] Czech Technical University – University Hospital Brno.
[3] Czech Gynaecological and Obstetrical Society.

The assumption for α_{ck}^{j} is that the evaluation for different c and k are independent on the observed data. This assumption is violated in practice since some examples are more difficult than others and each clinician has different level of expertise. The approach dealing with dependence on observed data was described in [24], however no significant improvements were acquired. Unlike [25] we formulate the model in a simplified way that is, every clinician provides one evaluation for each example. With this simplification we completely rule out the possible violation of conditional independence between two evaluations assigned by a clinician to a certain example.

The LCM considers clinical evaluation as a finite mixture of multinomial distributions. Finite mixture models [26] have fixed number of parameters and the standard method to estimate these parameters is expectation maximization (EM) algorithm [27]. Let us consider one particular set of evaluations $y^1, \ldots, y^J$. Then, it is assumed that these evaluations are from a mixture of initially specified C components in some unknown proportions $p_1, \ldots, p_C$. Each data point is a realization of the mixture probability mass function

$$p(y^1, \ldots, y^J|\boldsymbol{\theta}) = \sum_{c=1}^{C} p_c p(y^1, \ldots, y^J|\boldsymbol{\theta}_c), \tag{3}$$

where $\boldsymbol{\theta}$ include the unknown mixing proportion p_c (prevalence) and the elements of $\boldsymbol{\theta}_c$. Then, given a set of evaluations $\mathcal{D} = \{y_i^1, \ldots, y_i^J\}_{i=1}^N$ and vector of parameters $\boldsymbol{\theta} = \{\alpha_{ck}^j, p_c\}$, the likelihood corresponding to C component mixture is

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\sum_{c=1}^{C} p_c p(y_i^1, \ldots, y_i^J|\boldsymbol{\theta}_c) \right]. \tag{4}$$

We treat the unknown truth y_i as a latent (hidden) variable and use the EM algorithm to estimate it. We assume that $y_i^1, \ldots, y_i^J$ are independent (i.e. all clinicians make their evaluation independently) and that the evaluations are from multinomial distribution. Then, the likelihood function of the parameters $\boldsymbol{\theta}$ given $\mathcal{D}$ can be formulated as

$$p(y_i^1, \ldots, y_i^J|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\sum_{c=1}^{C} p_c \prod_{j=1}^{J} \prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j,k)} \right], \tag{5}$$

The maximum likelihood is found by maximizing the log likelihood function

$$\hat{\theta}_{\mathrm{ML}} = \{\alpha_{ck}^1, \ldots, \alpha_{ck}^J, p_1, \ldots, p_{C-1}\} = \arg\max_{\theta} \{\log p(\mathcal{D}|\boldsymbol{\theta})\}. \tag{6}$$

To maximize the log likelihood we followed the work of Dawid and Skene [25] and used the EM algorithm.

Estimation using the EM algorithm. The hidden variables to be estimated are multinomial parameter α_{ck}^j, prevalence of classes p_c, and true (unknown/hidden) evaluations $\{y_i\}_{i=1}^N$. If we would have known the true evaluation $\boldsymbol{y}$, the complete log-likelihood would be computed as

$$\log\ p(\mathcal{D}, \boldsymbol{y}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{c=1}^{C} \delta(y_i, c) \log \left[p_c \prod_{j=1}^{J} \prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j,k)} \right]. \tag{7}$$

In the **E-step** we compute the conditional expectation of y_i given the evaluations from clinicians $\mathcal{D}$ under the current estimates of parameters $\boldsymbol{\theta}$

$$\mathbb{E}\{\log p(\boldsymbol{y}|\mathcal{D}, \boldsymbol{\theta})\} = \sum_{i=1}^{N} \sum_{c=1}^{C} \mu_{ic} \log \left[p_c \prod_{j=1}^{J} \prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j,k)} \right], \tag{8}$$

where $\mu_{ic} = p(y_i = c|y_i^1, \ldots, y_i^J, \boldsymbol{\theta})$ is estimated probability of ground truth given the y_i^j and $\boldsymbol{\theta}$ and is proportional to

$$\mu_{ic} \propto p_c \prod_{j=1}^{J} \prod_{k=1}^{C} (\alpha_{ck}^j)^{\delta(y_i^j,k)}. \tag{9}$$

In the **M-step** we use the current estimates to maximize the conditional expectation. Taking a derivative of (8) equal to zero, the parameters α_{ck}^j and p_c are updated using the following equations

$$\alpha_{ck}^j = \frac{\sum_{i=1}^{N} \mu_{ic} \delta(y_i^j, k)}{\sum_{i=1}^{N} \mu_{ic}}, \quad p_c = \frac{1}{N} \sum_{i=1}^{N} \delta(\max_c(\mu_{ic}), c), \tag{10}$$

where $\max_c(\mu_{ic})$ assigns a class c that has the maximum probability. The E and M steps are repeated until convergence. The EM algorithm is guaranteed to converge to a local maximum only; therefore, it is usually restarted several times with different starting values. Another possible solution, used in this work, is to use the majority voting for initialization as proposed in [25]. The limit of log-likelihood convergence was set to 10^{-3}.

2.3.3. Latent class analysis with different number of classes

The latent model is powerful not only for estimating the latent class from multiple, possibly noisy, evaluations but could be also used to infer the number of classes the clinicians are actually using. Employing the LCA we can infer the number of classes, for which the evaluation would yield the best score – irrespective of the number of classes the clinicians used. In (**??**) we supposed a fixed number of classes. However, the guidelines are not precise, nor they are strictly followed by clinicians, leaving an open space for alternative evaluation. Our goal is to examine whether choosing different number of classes offers better description of clinical evaluation in terms of model fit. The extension to encompass different number of classes is straightforward. We replace C by a number R representing different number of latent classes

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\sum_{r=1}^{R} p_r \prod_{j=1}^{J} \prod_{k=1}^{C} (\alpha_{rk}^j)^{\delta(y_i^j,k)} \right], \tag{11}$$

where the same holds for α_{rk}^j as it did for α_{ck}^j. In our experiments we used the value of $R = \{2, 3, \ldots, 8\}$, obtaining models $M_2, M_3, \ldots, M_8$.

Number of estimated parameters. The number of estimated parameters ϑ increases rapidly with increasing R, J, and C and is computed as: $\vartheta = R - 1 + J \cdot [R(C - 1)]$. If the ϑ exceeds number of examples the model is not identifiable. The model is also not identifiable if the probabilities α_{rk}^j are sparse.

2.4. Rank of clinicians

The latent class model (LCM) can also be used to rank contribution of individual clinicians. The scoring/ranking based on detection of spammers was proposed by Raykar and Yu [21], where random evaluations were penalized. In our work, we had to adapt the penalization to reflect clinical evaluation, which we do not expect to be random. We use the following accuracy based score that is fairly simple and easily interpretable. Let $\mathbf{A}^j$ be a $C \times C$ confusion matrix with entries $[\mathbf{A}^j]_{ck} = \alpha^j_{ck}$. The diagonal elements represent probabilities of correct classifications with respect to latent class, $c = k$, and off-diagonal elements represent probabilities of misclassification, $c \neq k$. Consider the following confusion matrices for a good clinicians $\mathbf{A}^g$ and bad $\mathbf{A}^b$

$$\mathbf{A}^g = \begin{pmatrix} 0.9 & 0.1 & 0 \\ 0.1 & 0.9 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix} \quad \mathbf{A}^b = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.05 & 0.95 \\ 0 & 0.05 & 0.95 \end{pmatrix}. \tag{12}$$

The good clinicians $\mathbf{A}^g$ performed well on the first and second class and poorly on the third class, where the probability of correct decision dropped to 0.5. The bad clinician $\mathbf{A}^b$ correctly evaluated prevalently the third class. The accuracy based score for $C = K$ is defined as

$$\mathcal{S}^j_{acc} = \frac{1}{C}\left(\sum_{c=k} \mathbf{A}^j_{ck} - \sum_{c\neq k} \mathbf{A}^j_{ck}\right). \tag{13}$$

The score simply equals to summation of diagonal elements with subtraction of off-diagonal elements. In the case of matrices $\mathbf{A}^g$ and $\mathbf{A}^b$ the score yields $\mathcal{S}^g_{acc} = 0.53$ and $\mathcal{S}^b_{acc} = 0$, respectively. The score for the worst possible clinician is $\mathcal{S}_{acc} = -1$ and for the best possible is $\mathcal{S}_{acc} = 1$.

Ranking for different number of classes. The accuracy based score has a limitation in case the latent variable has different number of classes than the clinicians actually used. We focus only on the scenario when $R = 4$ because of the best model fit. Let $\mathbf{A}^j_{rk}$ be a matrix $[\mathbf{A}^j]_{rk} = \alpha^j_{rk}$, where $R \neq C$. When $R = 4$ and $C = 3$ the score is defined as

$$\mathcal{S}^j_{acc} = \frac{1}{R}\left(\sum_{r=k, r-1=k} \mathbf{A}^j_{rk} - \sum_{r\neq k, r-1\neq k} \mathbf{A}^j_{rk}\right). \tag{14}$$

The score allows misclassification over one latent class. The computation of the score is visualized in Table 1. We discuss the rationale for clinical evaluation later. Note that the score serves for ranking clinicians for particular choice of r. We detail the comparison of the models (i.e. their fit) for various r in the next section.

Table 1: Score computation for the four latent classes ($R = 4$). The latent classes are in rows and classes given by clinicians in columns. Legend: the correct classification is marked by (+) and incorrect by (−).

	$k=1$	$k=2$	$k=3$
$r=1$	+	−	−
$r=2$	+	+	−
$r=3$	−	+	+
$r=4$	−	−	+

2.5. Model selection and fit

We can use various techniques to evaluate a model fit and to determine which model is more appropriate for different values of R. Increasing R from two to eight increases the model fit but also increases the possibility of over-fitting. Additionally, higher R values lead to estimation of more model parameters. A trade-off between better model fit and number of parameters to be estimated is usually sought and tackled by penalizing the log likelihood using a function of parameters $\boldsymbol{\theta}$ that are to be estimated. The two most common measures are the Akaike information criterion (AIC) [28] and Bayes information criterion (BIC) [29]. For a likelihood L the AIC and BIC are defined as

$$AIC(r) = -2\ln L + 2\vartheta,$$
$$BIC(r) = -2\ln L + \vartheta \ln N.$$

The better is the model the lower BIC and/or AIC measures are obtained. Usually the AIC overestimates the number of R while BIC underestimates it. Therefore a compromise between these is often sought.

2.6. Stability of clinical evaluation

We use majority voting for the description of stability, but any other method for aggregation could be used. The motivation for analysis of the stability of clinical evaluation follows: Let us consider majority voting of J clinicians. It would be interesting to know whether the created majority was obtained simply by a chance or whether the majority is stable and the possible variability in clinicians' decision was cancelled out by using a sufficient number of clinicians. We summarize the definition of stability in Proposition 1.

Proposition 1. *We consider a majority vote of J clinicians stable if a majority voting of J + 1 clinicians is not different (measured by proportion of agreement).*

Algorithm 1: Stability of clinical evaluation

Input: $Q = \{3, 4, \ldots, 8\}$ number of clinicians, $\mathbf{Y}$ clinical evaluation of size $N \times J$, mv_J majority vote of all J clinicians, mv_b majority vote of combination b of clinicians
Result: pa - proportion of agreement
begin
  for $q \in Q$ **do**
    $comb \leftarrow \binom{J}{q}$ – all combinations of q clin. from J
    for $b \in comb$ **do**
      $\mathbf{Y}_b = \mathbf{Y}(:, b)$ – get evaluation of selected b
      $mv_b \leftarrow majorityVoting(\mathbf{Y}_b)$
      $pa(q, b) \leftarrow proportionOfAgreement(mv_b, mv_J)$
    end
  end
end

In the proposition, the term "different", our criterion, is not rigorous and allows various approaches to be used, i.e. statistical testing. However, the statistical evaluation is not that straightforward as the created majority votes are not independent. In

order to analyze the stability we performed the following experiment: we computed majority votes (MV) for all combinations of clinicians $\binom{J}{q}$, where $q \in Q; Q = \{3, 4, \ldots, J-1\}$. Then we compared the majority obtained with the majority vote of all clinicians, $J = 9$. The procedure is shown in Algorithm 1.

3. Results and discussion

3.1. Stability of evaluation in majority voting and latent class model

The stability of clinical evaluation in majority voting (MV) and latent class model (LCM), irrespective of evaluated classes, is presented in Figure 1. For both, MV and LCM, stability increases with increasing number of clinicians in ensemble. The stability for LCM is better with higher number of clinicians while for the lower number the MV performs better with lower variance. Thus we conclude that for $q \leq 5$ the MV should be preferred and for $q > 5$ the LCM should be favoured. The same conclusion holds when the overall evaluation is split into the individual classes as shown in Figure 2. Especially for the pathological class with $q > 5$ the LCM provides more stable aggregation than the MV. These conclusions are confirmed with statistical testing[4] for differences between models MV and LCM for different combinations of clinicians. The statistical significance, $p < 0.05$, is marked in Figures 1 and 2 with an asterisk. Note, that in some cases (for even number clinicians) the majority vote had to be determined by flip of a fair coin since the votes were equal. This phenomenon can explain larger improvements from even q to odd q, e.g. in Figure 1 the improvement of the MV model's proportion of agreement from $q = 4$ to $q = 5$ in comparison to $q = 5$ to $q = 6$.

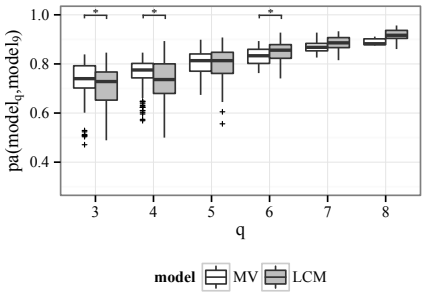


Figure 1: Stability of majority voting (MV) and latent class model (LCM) for all classes and different combinations of clinicians. Legend: pa(model$_q$,model$_9$) stands for proportion of agreement between model for q clinicians and model for nine clinicians, where the model is replaced either by MV or LCM (i.e. it leads to pa(mv$_q$,mv$_9$) and pa(lcm$_q$,lcm$_9$)), *marks statistical significance on $p < 0.05$.

3.2. Latent class analysis – different number of classes

We analyzed the clinical evaluation using the latent class model for different number of classes. The model fit statistics are shown in Table 2. The progression of AIC and BIC for the increasing value of r is shown in Figure 3. Clinicians evaluate the CTG into three FIGO classes (normal, suspicious, and pathological). Nevertheless, from Figure 3 we can conclude that the best fit is obtained for the model M_4. From the model M_3 to M_4 both measures, AIC and BIC, decreases. The BIC starts rising from M_4 to M_5 while the AIC only slightly decreases, hence the best fitted model is M_4.

Table 2: Fit statistics for different number of classes (df – degrees of freedom, AIC – Akaike information criterion, BIC – Bayes information criterion). The lower the AIC and/or BIC, the more fit is the model.

model	df	AIC	BIC
M_2	515	7316	7476
M_3	496	6842	7083
M_4	477	6677	7000
M_5	458	6656	7062
M_6	439	6666	7154
M_7	420	6669	7239
M_8	401	6688	7340

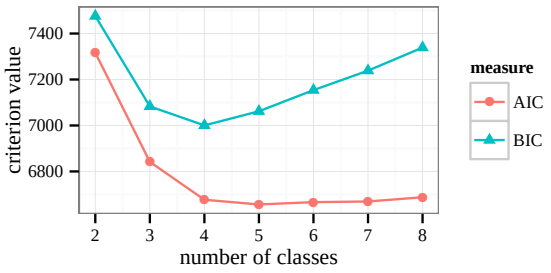
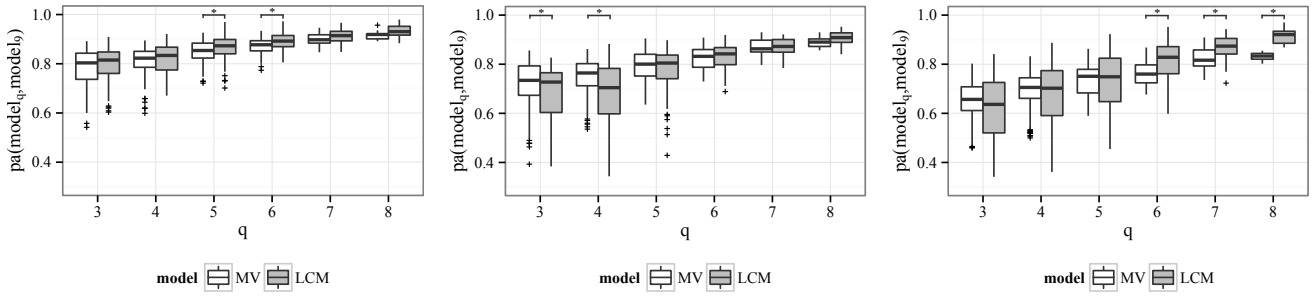


Figure 3: Progression of AIC and BIC for different number of classes ($r = \{2, 3 \ldots, 8\}$).

In order to have better insight into models M_3 and M_4 we evaluated the multinomial parameter α_{ck}^{j} (conditional probability of assigning class k with respect to the latent class c). Models M_3 and M_4 are presented in Figures 4 and 5 respectively. The clinicians are marked with numbers $(1, 2, \ldots, 9)$ and their evaluation are separated with respect to estimated latent class. Note that prevalence of evaluated classes of each clinician could be easily observed in Figure 4. Prevalence of k-th class for the j-th clinician is determined as $p_k^j = (1/C) \sum_{c=1}^{C} \alpha_{ck}^{j}$. For instance, for the first clinicians, $j = 1$, we compute $p_1^1 = (1/C) \sum_{c=1}^{C} \alpha_{c1}^{1} = (1/3)(1 + 0.49 + 0.03) = 0.51$, $p_2^1 = 0.41$, and $p_3^1 = 0.1$.

For the M_3 the latent class can be separated into normal ($c = 1$), suspicious ($c = 2$), and pathological ($c = 3$) based on the majority of clinical evaluation. For the M_4 the situation is more complicated. The assignment of classes is rather intuitive and could be determined with help of knowledge about proportions of classes of each clinician. For the first class of the M_4 model, that was ex-post labelled as normal ($r = 1$), we can see in Figure 5 that majority of clinicians' evaluation was normal. For the second class, ex-post labelled to normal/suspicious ($r = 2$) we can observe discrepancy in evaluation of different clinicians. Prevalently normal evaluation by clinicians 1,

[4] Wilcoxon rank-sum test for paired samples of not normally distributed data was used (normality tested using Kolmogorov–Smirnov test with Lillefors correction); $p < 0.05$ was considered as significant.



(a) normal (b) suspicious (c) pathological

Figure 2: Stability of majority voting and latent class model (LCM) with respect to different evaluation: normal, suspicious, and pathological. pa(model$_q$,model$_9$) stands for proportion of agreement between model for q clinicians and model for nine clinicians, where the model is replaced either by MV or LCM (i.e. it leads to pa(mv$_q$,mv$_9$) and pa(lcm$_q$,lcm$_9$)), *marks statistical significance on $p < 0.05$.

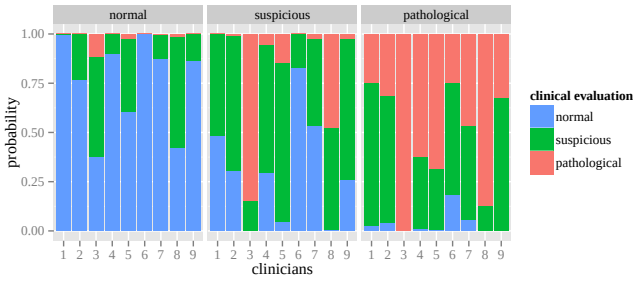


Figure 4: Conditional probability of clinical evaluation to the latent class. Model M_3. Estimated latent classes were as follows: normal, suspicious, and pathological (shown in grey headings). Latent class prevalences: $P(normal) = 0.30$, $P(suspicious) = 0.45$, $P(pathological) = 0.25$.

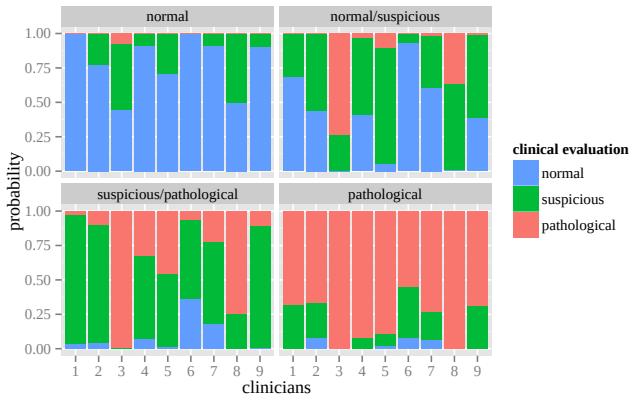


Figure 5: Conditional probability of clinical evaluation to the latent class. Model M_4. Estimated latent classes were named as follows: normal, normal/suspicious, suspicious/pathological, and pathological (shown in grey headings). Latent class prevalences: $P(normal) = 0.25$, $P(normal/suspicious) = 0.38$, $P(suspicious/pathological) = 0.29$, $P(pathological) = 0.08$.

6, and 7, prevalently suspicious by clinicians 2, 4, 5, 8, and 9, and prevalently pathological by clinician 3. Considering proportions of evaluation, we can see that clinician 6 mostly evaluated CTG as normal, while clinician 3 mostly evaluated CTG as pathological. In this case we can neglect them for decision on class label and use the other clinicians' votes. Hence we labelled it as normal/suspicious. The other classes were: suspicious/pathological ($r = 3$), and pathological ($r = 4$). Again, we would like to note that the label of final class is based on the intuition rather than rigorous classification.

Rank of clinicians. The score of clinicians for M_3 and M_4 was determined using (13) and (14) respectively. The confusion matrix in Table 3 details the computation of (14).

The rank of clinicians for majority voting (first iteration of the M_3) and the M_3 model are presented in Table 4; the ranks for the M_4 are shown in Table 5. Note that for comparison of models with different classes we are interested in rank of clinicians rather than absolute values of score since the scores are generally higher for the M_4 because more elements of confusion matrix in Table 3 are considered as correct. The workplace was anonymised by assignment of letters A–F. There are three pairs of clinicians coming from the same workplace (B, E, and F). Because of low number of clinicians coming from the same workplace the difference between workplaces was not evaluated. Experience of clinicians is not presented because it would compromise their anonymity. Nevertheless we found no relationship between the score and the experience.

The rank of clinicians for majority voting and model M_3 are essentially the same. The ranks for model M_4 in comparison to majority voting and M_3 are different in the way that clinicians 1, 5, and 8 change their ranks markedly. The reasons were as follows: clinician 1 improved on pathological recordings. In the model M_3 this clinician mixed suspicious evaluation into the pathological class ($c = 3$) while for the M_4 the proportion of suspicious in pathological class ($r = 4$) was significantly lowered. Clinician 5 was ranked lower mainly because of mixing suspicious evaluation into normal class ($r = 1$). The drop in rank for clinician 8 in M_4 with regard to M_3 is due to excellent performance on pathological class ($c = 3$), for the M_4 his evaluation were mixed into different latent classes; mainly suspicious evaluation to normal class ($r = 1$).

4. Discussion

Agreement on CTG evaluation. The latent class analysis provides broader insight into obstetricians' decisions on CTG evaluation. It was shown that when clinicians evaluate CTG independently the agreement using majority voting is hard to reach.

Table 3: Score computation for the model M_4. The latent classes are in rows and classes given by clinicians in columns. Legend: the correct classification is marked by (+) and incorrect by (−), i.e. when a clinician assigned normal class then latent classes normal and normal/suspicious were considered as correct.

	normal ($k = 1$)	suspicious ($k = 2$)	pathological ($k = 3$)
normal ($r = 1$)	+	−	−
normal/suspicious ($r = 2$)	+	+	−
suspicious/pathological ($r = 3$)	−	+	+
pathological ($r = 4$)	−	−	+

Table 4: Score and rank of individual clinicians for majority voting and model M_3. (#clin. – number of a clinician, WP – workplace).

	majority voting			M_3		
rank	$\mathcal{S}_{acc}$	#clin.	WP	$\mathcal{S}_{acc}$	clin.	WP
1	0.1778	4	B	0.4487	4	B
2	0.1202	5	E	0.3984	5	E
3	0.1178	9	E	0.2677	9	E
4	0.1111	7	C	0.2061	8	F
5	0.1029	1	B	0.1907	7	C
6	0.0833	8	F	0.1753	2	A
7	0.0736	2	A	0.1716	1	B
8	0.0060	6	D	0.0202	3	F
9	−0.0313	3	F	−0.0556	6	D

Table 5: Score and rank of individual clinicians for model M_4. (#clin. – number of a clinician, WP – workplace)

rank	$\mathcal{S}_{acc}$	#clin.	WP
1	0.8431	4	B
2	0.8202	1	B
3	0.7595	9	E
4	0.6976	7	C
5	0.6668	5	E
6	0.6395	2	A
7	0.5916	6	D
8	0.4379	8	F
9	0.2173	3	F

The main reason for difficulties in reaching the agreement is the inherent large inter-observer variability that remains problematic even when using evaluations acquired from large number of clinicians.

We have shown that in the case when seven clinicians have evaluated (by majority voting) a CTG record as pathological, an additional clinician might change the current majority vote in about 17% of cases. We have shown that to weight clinical decisions is superior to the majority voting, and it is in fact more natural way to reach an agreement in the experience based field.

We employed novel approach using the latent class modelling where the latent (hidden) true class of clinical evaluation was estimated iteratively by changing weights of individual clinicians (i.e. their reliability in the given ensemble of clinicians). We have shown that such an approach leads to better stability of evaluation where a large improvement has been obtained especially for the pathological evaluation.

Varying number of latent classes. The latent class model (LCM) is powerful tool not only for estimation of a latent class from multiple, possibly noisy, evaluations but could be also used to infer the number of classes the clinicians are actually using – although possibly unconsciously. By investigating the LCM model with varying number of classes, we tried to attribute to the ongoing discussion of how many classes should be used for CTG evaluation. We found that the model has the best fit for four classes. The main advantage of using four instead of three classes is in better separation of pathological records from the other ones. In other words, for the M_4 model there is a clear group of pathological records, for which there is a good agreement among clinicians; for the other classes the evaluation is more diverse and splitting these classes further did neither contribute to better model fit, nor did it lower clinician's variability.

Rank of clinicians. For the latent class model we assessed the contribution of each clinicians using a scoring function. Interestingly the ranks of clinicians for the majority voting and model M_3 were essentially the same. This implies that the model M_3 is in fact data-driven weighted majority voting that gives more stable results, cf. Section 3.1. We have found no relationship between clinical experience and rank of clinicians.

We have shown that latent class analysis is more suitable approach than majority voting. However its limitation is the need to re-learn the model when a new evaluation is obtained. The generalization of the achieved results to the whole population is always difficult in agreement analysis. The database was selected without knowledge about any specific type of evaluation hence we believe that results are valid also with regard to the whole population. In terms of number of evaluated records and number of clinicians taking part in annotation, our study is the largest study that has been performed so far. The typical size of other studies was 30-50 recordings [23, 30].

5. Conclusion

In this paper we have described a novel approach for analysis of CTG evaluation – the latent class model (LCM). With the LCM model with varying number of classes we have contributed to the discussion on how many classes should be used for CTG evaluation. We have shown that the model has the best fit for 4-tier classification. The difference between 3 and 4 classes lies in better separation of pathological records from the other ones. In other words, there is a clear pathological group for which there is a good agreement among clinicians. We have proved that even with a high number of clinicians the agreement (majority voting) cannot be reached. The lack of agreement can be contributed to the large inter- and intra-observer variability. The latent class model allowed to examine agreement in more detail and provided more stable aggregation of clinical evaluation. A large improvement have been obtained especially for the pathological evaluation.

The goal of the paper was to use data-driven approach to agreement analysis and not to change the current practice in the obstetrics ward. The presented results support the arguments

asking for modernization of the FIGO guidelines. The existing and widely reported inter- and intra-observer variability is major suspect in the difficulties to establish agreement among clinicians. If we exclude consensus achieved in a panel discussion – often based on hospital hierarchy rather than objective facts – the guidelines do not provide stable basis for easy agreement. Four evaluation classes used unconsciously by our group of clinicians, as revealed by the latent class analysis, suggest that the definitions of the classes in FIGO guidelines are ambiguous and difficult to understand. Such a conclusion is not new but in this paper it is supported by the data-driven analysis.

In the future there is a need for even larger pool of clinicians in order to achieve 100% stable agreement on a CTG record. We plan to involve more clinicians to evaluate the CTU-UHB database[5] in order to confirm the established relationships presented in Figures 1 and 2. In addition, we plan to further examine the pathological group where there is a good agreement (model M_4).

Acknowledgements

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. We would like to acknowledge all medical experts that participated on the annotation: A. Hudec, V. Korečko, M. Kacerovský, M. Koucký, M. Procházka, J. Segeťa, and O. Šimetka. We are grateful to Philips Healthcare for providing a tool to convert data from the OB TraceVue and local representative Monika Jiráčková from S&T for providing us with her expertise.

References

[1] T. P. Sartwelle, Electronic fetal monitoring: a bridge too far., J Leg Med 33 (3) (2012) 313–379.

[2] FIGO, Guidelines for the Use of Fetal Monitoring, International Journal of Gynecology & Obstetrics 25 (1986) 159–167.

[3] S. C. Blackwell, W. A. Grobman, L. Antoniewicz, M. Hutchinson, C. Gyamfi Bannerman, Interobserver and intraobserver reliability of the NICHD 3-Tier Fetal Heart Rate Interpretation System., Am J Obstet Gynecol 205 (4) (2011) 378.e1–378.e5.

[4] RCOG, Royal College of Obstetricians and Gynaecologists. The use of electronic fetal monitoring. Evidence-based clinical guidelines, RCOG Press, London (2001).

[5] G. A. Macones, G. D. V. Hankins, C. Y. Spong, J. Hauth, T. Moore, The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines., J Obstet Gynecol Neonatal Nurs 37 (5) (2008) 510–515.

[6] ACOG, American College of Obstetricians and Gynecologists Practice Bulletin No. 106: Intrapartum fetal heart rate monitoring: nomenclature, interpretation, and general management principles., Obstet Gynecol 114 (1) (2009) 192–202.

[7] D. A. de Campos, J. Bernardes, 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) (2010) 1 – 6.

[8] B. S. Schifrin, The CTG and the timing and mechanism of fetal neurological injuries., Best Pract Res Clin Obstet Gynaecol 18 (3) (2004) 437–456.

[5] We would welcome any researcher/clinicians interested in CTG annotation.

[9] J. T. Parer, T. Ikeda, A framework for standardized management of intrapartum fetal heart rate patterns., Am J Obstet Gynecol 197 (1) (2007) 26.e1–26.e6.

[10] J. T. Parer, T. Ikeda, T. L. King, The 2008 National Institute of Child Health and Human Development report on fetal heart rate monitoring., Obstet Gynecol 114 (1) (2009) 136–138.

[11] J. Parer, E. Hamilton, Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation, American Journal of Obstetrics and Gynecology 203 (5) (2010) 451.e1–451.e7.

[12] M. D. Tommaso, V. Seravalli, A. Cordisco, G. Consorti, F. Mecacci, F. Rizzello, Comparison of five classification systems for interpreting electronic fetal monitoring in predicting neonatal status at birth., J Matern Fetal Neonatal Med 26 (5) (2013) 487–490.

[13] J. Coletta, E. Murphy, Z. Rubeo, C. Gyamfi-Bannerman, The 5-tier system of assessing fetal heart rate tracings is superior to the 3-tier system in identifying fetal acidemia., Am J Obstet Gynecol 206 (3) (2012) 226.e1–226.e5.

[14] D. A. Miller, L. A. Miller, Three-tier versus five-tier fetal heart rate classification systems., Am J Obstet Gynecol 207 (6) (2012) e8–9; author reply e9.

[15] C. C. Santos, A. C. Pereira, J. Bernardes, Agreement studies in obstetrics and gynaecology: inappropriateness, controversies and consequences., BJOG 112 (5) (2005) 667–669.

[16] A. R. Feinstein, D. V. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes, Journal of clinical epidemiology 43 (6) (1990) 543–549.

[17] D. V. Cicchetti, A. R. Feinstein, High agreement but low kappa: II. Resolving the paradoxes, Journal of clinical epidemiology 43 (6) (1990) 551–558.

[18] L. Zach, V. Chudáček, M. Huptych, J. Spilka, M. Burša, L. Lhotská, 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, Beijing, China, Springer, 2013, pp. 1280–1282.

[19] V. Chudáček, J. Spilka, M. Burša, P. Janků, L. Hruban, M. Huptych, L. Lhotská, Open access intrapartum CTG database., BMC Pregnancy Childbirth 14 (1) (2014) 16.

[20] P. J. Boland, Majority systems and the Condorcet jury theorem, The Statistician 38 (3) (1989) 181–189.

[21] V. C. Raykar, S. Yu, Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks, Journal of Machine Learning Research 13 (2012) 491–518.

[22] E. Blix, O. Sviggum, K. S. Koss, P. Oian, Inter-observer variation in assessment of 845 labour admission tests: comparison between midwives and obstetricians in the clinical setting and two experts., BJOG 110 (1) (2003) 1–5.

[23] C. Vayssiere, V. Tsatsaris, O. Pirrello, C. Cristini, C. Arnaud, F. Goffinet, Inter-observer agreement in clinical decision-making for abnormal cardiotocogram (CTG) during labour: a comparison between CTG and CTG plus STAN., BJOG 116 (8) (2009) 1081–7; discussion 1087–8.

[24] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, P. Malvern, Modeling annotator expertise: Learning when everybody knows a bit of something, in: International Conference on Artificial Intelligence and Statistics, Vol. 9, 2010, pp. 932–939.

[25] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the EM algorithm, Applied Statistics 28 (1979) 20–28.

[26] G. McLachlan, D. Peel, Finite Mixture Models, New York, John Wiley & Sons., 2000.

[27] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological) (1977) 1–38.

[28] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov, F. Csaki (Eds.), Second International Symposium on Information Theory, Akadémiai Kiado, Budapest, 1973, pp. 267–281.

[29] G. Schwarz, Estimating the Dimension of a Model, The Annals of Statistics 6 (2) (1978) 461–464.

[30] R. D. Keith, S. Beckley, J. M. Garibaldi, J. A. Westgate, E. C. Ifeachor, K. R. Greene, A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram., Br J Obstet Gynaecol 102 (9) (1995) 688–700.