Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

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

ARTICLE INFO

Article history:
Received 4 October 2013
Accepted 7 April 2014
Available online xxxx

Keywords:
Cardiotocography
Fetal heart rate
Observer variation
Biomedical informatics
Decision making
Latent class analysis

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.

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 [4–6]. 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.

* Corresponding author. Address: Karlovo namesti 13, 121 35 Prague 2, Czech Republic. Tel.: +420 224 357 325.
E-mail address: spilka.jiri@fel.cvut.cz (J. Spilka).

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

http://dx.doi.org/10.1016/j.jbi.2014.04.010
1532-0464/

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 measured. The kappa coefficient, proportion of agreement, and intra-class 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] Czech Technical University – University Hospital Brno.
[3] Czech Gynaecological and Obstetrical Society.

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

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

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