Archives of Gynecology and Obstetrics (2019) 300:7–14
https://doi.org/10.1007/s00404-019-05151-7

REVIEW

Use of artificial intelligence (AI) in the interpretation of intrapartum fetal heart rate (FHR) tracings: a systematic review and meta-analysis

Jacques Balayla[1] · **Guy Shrem**[1]

Received: 12 January 2019 / Accepted: 4 April 2019 / Published online: 3 May 2019

Abstract

Objectives To determine the degree of inter-rater reliability (IRR) between human and artificial intelligence (AI) interpretation of fetal heart rate tracings (FHR), and to determine whether AI-assisted electronic fetal monitoring interpretation improves neonatal outcomes amongst laboring women.

Data sources We searched Medline, EMBASE, Google Scholar, Scopus, ISI Web of Science and Cochrane database search, as well as PubMed (www.pubmed.gov) and RCT registry (www.clinicaltrials.gov) until the end of October 2018 to conduct a systematic review and meta-analysis comparing visual and AI interpretation of EFM in labor. Similarly, we sought out all studies evaluating the IRR between AI and expert interpretation of EFM.

Tabulation, integration and results Weighed mean Cohen's Kappa was calculated to assess the global IRR. Risk of bias was assessed using the Cochrane Handbook for Systematic Reviews of Interventions. We used relative risks (RR) and a random effects (RE) model to calculate weighted estimates. Statistical homogeneity was checked by the χ^2 test and I^2 using Review Manager 5.3.5 (The Cochrane Collaboration, 2014.) We obtained 201 records, of which 9 met inclusion criteria. Three RCT's were used to compare the neonatal outcomes and 6 cohort studies were used to establish the degree of IRR between both approaches of EFM evaluation. With regards to the neonatal outcomes, a total of 55,064 patients were included in the analysis. Relative to the use of clinical (visual) evaluation of the FHR, the use of AI did not change the incidence rates of neonatal acidosis, cord pH below < 7.20, 5-min APGAR scores < 7, mode of delivery, NICU admission, neonatal seizures, or perinatal death. With regards to the degrees of inter-rater reliability, a weighed mean Cohen's Kappa of 0.49 [0.32–0.66] indicates moderate agreement between expert observers and computerized systems.

Conclusion The use of AI and computer analysis for the interpretation of EFM during labor does not improve neonatal outcomes. Inter-rater reliability between experts and computer systems is moderate at best. Future studies should aim at further elucidating these findings.

Keywords Artificial intelligence · Computer · Fetal monitoring · Fetal heart rate · Inter-rater reliability · Neonatal outcomes

Introduction

Artificial intelligence (AI) is increasingly finding applications in clinical medicine, particularly, in areas requiring multiple input assessment and rapid decision-making [1, 2]. One such area is the electronic fetal monitoring (EFM) of women in labor, where despite the existence of standardized guidelines, considerable inter and intra-observer variability in the interpretation of intrapartum fetal heart rate (FHR) tracings remains [3, 4]. The FHR pattern observed in labor reflects the fetal cardiac and central nervous system responses to changes in blood pressure, blood gases, and acid–base status, and serves thus as a marker of fetal wellbeing [5]. The rationale for intrapartum FHR monitoring is that identification of FHR changes potentially associated with inadequate fetal oxygenation may enable timely intervention to reduce the likelihood of hypoxic injury or death [6].

In this clinical scenario, the theoretical benefit of using AI-assistance lies in reducing inter-observer variability as well as in providing real-time, beat-to-beat interpretation of the FHR and fetal wellbeing to ensure that opportunities for intervention and injury prevention are not missed. Similarly, it is conceivable that by providing a more consistent

✉ Jacques Balayla
jacques.balayla@mail.mcgill.ca

[1] Department of Obstetrics and Gynecology, McGill University, Montreal, QC, Canada

interpretation of EFMs, AI-assistance may have implications in medico-legal cases revolving around FHR interpretations. That said, while the use of continuous EFM to reduce neonatal morbidity is debated in its own right [7], whether neonatal outcomes are improved with the use of AI has only recently been studied and remains controversial [8–10].

In this study, we conducted a systematic review and meta-analysis to accomplish two objectives: (1) to determine the degree of inter-rater reliability (IRR) between human and AI interpretation of FHR tracings in labor, and (2) to determine whether AI-assisted EFM interpretation in labor improves neonatal outcomes.

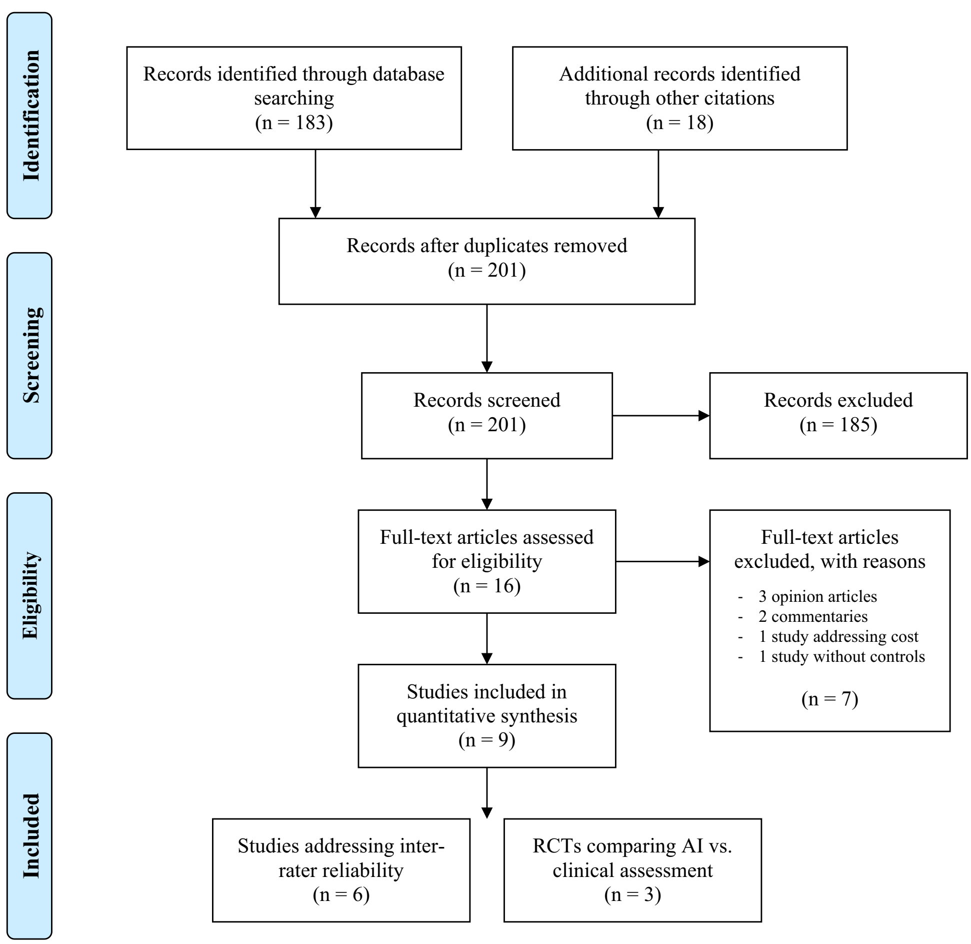
Sources

We performed a Medline, EMBASE, Google Scholar, Scopus, ISI Web of Science and Cochrane database search, as well as PubMed (www.pubmed.gov) and RCT registry (www.clinicaltrials.gov) search until the end of October 2018 from the last 20 years using the following Boolean search criteria: '[(Artificial Intelligence OR AI) OR (computer analysis OR computer evaluation) AND (fetal heart rate OR FHR) OR (electronic fetal monitoring OR EFM) OR (cardiotocography)]. Other than restricting the search to human studies and to articles in English, the only limiting categorical term used was the restriction of studies in the intrapartum period. The reference lists and bibliographies of included studies were then searched for other salient and pertinent manuscripts. Finally, manual searches of studies belonging to research teams having prior publications on AI and computer-evaluated tracings were undertaken, and other pertinent studies retrieved. This review was modeled on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement, and search flowchart depicting the search strategy is illustrated in Fig. 1.

Study selection

Both authors independently examined the electronic search results for reports of possibly relevant studies and those reports were retrieved and analyzed in further detail. Published RCTs were eligible for inclusion if they compared AI-assisted with visual (human) interpretation of the EFM during labor, whereas observational studies

Fig. 1 PRISMA flowchart for the systematic review of studies



were eligible if they reported on the degree of inter-rater reliability, irrespective of neonatal outcomes. All studies were assessed following pre-determined quality criteria. Validity of RCTs was assessed in terms of method of randomization, presence of a power calculation, unit of analysis used, use of an intention-to-treat analysis and risk of bias. We did not find the need to contact any corresponding author of a report in an attempt to retrieve missing data given that data to carry out a meta-analysis was complete. We made no distinction between the type of AI software used, the length of the tracing portion evaluated, the number of expert observers, nor the year or country of origin of the study in question.

Risk of bias was assessed using the Cochrane Handbook for Systematic Reviews of Interventions (results not shown). We used relative risks (RR) and a random effects (RE) model with the Mantel–Haenszel method to calculate weighted estimates and their 95% confidence intervals (CI) where appropriate. Forest plots are provided for visualization of the results. Statistical heterogeneity between results of studies was examined by inspecting the scatter in the data points on the graphs and the overlap of CIs, and by checking the χ^2 and I^2 statistics. For the pooled calculation of the degree of IRR, we used the method developed by Shuyan Sun to calculate the weighed mean Cohen's Kappa in meta-analytic form [11]. The Review Manager 5.3.5 software (Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration 2014) was used to combine data for the meta-analysis. This meta-analysis was exempt from institutional review board (IRB) approval because of the nature of the research design (review article), as well as the lack of use of identified patient data.

Results

We obtained 201 records, of which 9 met inclusion criteria (Table 1). Three RCT's were used to compare the neonatal outcomes and 6 cohort studies were used to establish the degree of IRR between both approaches. With regards to the neonatal outcomes, a total of 55,064 patients were included in the analysis. Relative to the use of clinical (visual) evaluation of the FHR, the use of AI did not change the incidence rates of neonatal acidosis (RR [95% CI]): 0.72 [0.37, 1.40], cord pH below <7.20: 0.90 [0.77, 1.05], 5-min APGAR scores <7: 0.73 [0.45, 1.19], mode of delivery, NICU admission: 0.86 [0.66, 1.11], neonatal seizures: 0.92 [0.60, 1.39], or perinatal death: 1.28 [0.46, 3.61] ($p>0.05$) (Fig. 2). With regards to the degrees of inter-rater reliability, a weighed mean Cohen's Kappa of 0.49 [0.32–0.66] indicates moderate agreement between expert evaluation and AI systems (Table 2).

Discussion

Previous studies have suggested that a key element in substandard intrapartum care is the failure of clinicians to recognize an abnormal fetal heart-rate pattern [8]. In this study, we sought to determine whether AI-assisted evaluation of EFM was a reliable method of FHR interpretation and whether its implementation improved neonatal outcomes. Our findings suggest that as of yet, AI interpretation does not appear to improve outcomes and has yet to prove its reliability relative to expert observers when evaluating the FHR.

The development of a systematic and objective method of analysis and diagnosis in EFM is highly desirable, and indeed, there have been several attempts to develop computerized systems for a quantitative/qualitative analysis of

Table 1 Characteristics of studies included in the meta-analysis

	Ignatov et al. [9]	Nunes et al. [10]	The INFANT Trial [8]
Year	2016	2017	2017
Country	Bulgaria	U.K.	U.K + Ireland
Study design	RCT	RCT	RCT
Number of patients	720	7730	46,042
AI technology	Nexus-obstetrics	Omniview-SisPorto	Infant-K2
Inclusion criteria	>18 years of age Singleton pregnancy Cephalic position No known fetal structural abnormalities	16 years of age Singleton pregnancy Cephalic presentation 36 completed weeks of gestation or greater No known major fetal malformations In active labor but not in active second stage	>16 years of age Singleton or twin pregnancy 35 completed weeks of gestation or greater No known major malformations

Study or Subgroup	AI Events	AI Total	Clinical Events	Clinical Total	Weight	Risk Ratio M-H, Random, 95% CI
1.1.1 Acidosis						
Ignatov et al.	5	360	16	360	22.9%	0.31 [0.12, 0.84]
Nunes et al.	16	3961	22	3769	32.7%	0.69 [0.36, 1.32]
The INFANT Trial	148	23263	131	23351	44.4%	1.13 [0.90, 1.43]
Subtotal (95% CI)		**27584**		**27480**	**100.0%**	**0.72 [0.37, 1.40]**
Total events	169		169			
Heterogeneity: $Tau^2 = 0.24$; $Chi^2 = 7.64$, df = 2 (P = 0.02); $I^2 = 74\%$						
Test for overall effect: Z = 0.98 (P = 0.33)						
1.1.2 Blood cord pH < 7.15–7.20						
Ignatov et al.	25	360	47	360	9.7%	0.53 [0.33, 0.84]
Nunes et al.	367	3961	372	3769	39.3%	0.94 [0.82, 1.08]
The INFANT Trial	1625	23263	1695	23351	51.0%	0.96 [0.90, 1.03]
Subtotal (95% CI)		**27584**		**27480**	**100.0%**	**0.90 [0.77, 1.05]**
Total events	2017		2114			
Heterogeneity: $Tau^2 = 0.01$; $Chi^2 = 6.21$, df = 2 (P = 0.04); $I^2 = 68\%$						
Test for overall effect: Z = 1.31 (P = 0.19)						
1.1.3 5-min APGAR <7						
Ignatov et al.	9	360	18	360	29.4%	0.50 [0.23, 1.10]
Nunes et al.	47	3961	52	3769	70.6%	0.86 [0.58, 1.27]
Subtotal (95% CI)		**4321**		**4129**	**100.0%**	**0.73 [0.45, 1.19]**
Total events	56		70			
Heterogeneity: $Tau^2 = 0.05$; $Chi^2 = 1.46$, df = 1 (P = 0.23); $I^2 = 32\%$						
Test for overall effect: Z = 1.26 (P = 0.21)						
1.1.4 Cesarean Delivery						
Ignatov et al.	50	360	81	360	12.5%	0.62 [0.45, 0.85]
Nunes et al.	809	3961	722	3769	39.9%	1.07 [0.97, 1.17]
The INFANT Trial	5669	23263	5555	23351	47.6%	1.02 [0.99, 1.06]
Subtotal (95% CI)		**27584**		**27480**	**100.0%**	**0.98 [0.86, 1.11]**
Total events	6528		6358			
Heterogeneity: $Tau^2 = 0.01$; $Chi^2 = 10.34$, df = 2 (P = 0.006); $I^2 = 81\%$						
Test for overall effect: Z = 0.35 (P = 0.73)						
1.1.5 Instrumental Delivery						
Ignatov et al.	2	360	4	360	0.2%	0.50 [0.09, 2.71]
Nunes et al.	1252	3961	1112	3769	41.4%	1.07 [1.00, 1.15]
The INFANT Trial	5698	23263	5765	23351	58.4%	0.99 [0.96, 1.02]
Subtotal (95% CI)		**27584**		**27480**	**100.0%**	**1.02 [0.95, 1.10]**
Total events	6952		6881			
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 4.75$, df = 2 (P = 0.09); $I^2 = 58\%$						
Test for overall effect: Z = 0.63 (P = 0.53)						
1.1.6 NICU Admission						
Ignatov et al.	7	360	21	360	8.0%	0.33 [0.14, 0.77]
Nunes et al.	137	3961	143	3769	39.5%	0.91 [0.72, 1.15]
The INFANT Trial	653	23263	690	23351	52.5%	0.95 [0.85, 1.06]
Subtotal (95% CI)		**27584**		**27480**	**100.0%**	**0.86 [0.66, 1.11]**
Total events	797		854			
Heterogeneity: $Tau^2 = 0.03$; $Chi^2 = 5.88$, df = 2 (P = 0.05); $I^2 = 66\%$						
Test for overall effect: Z = 1.15 (P = 0.25)						
1.1.7 Neonatal Seizures						
Ignatov et al.	3	360	5	360	8.6%	0.60 [0.14, 2.49]
The INFANT Trial	39	23263	41	23351	91.4%	0.95 [0.62, 1.48]
Subtotal (95% CI)		**23623**		**23711**	**100.0%**	**0.92 [0.60, 1.39]**
Total events	42		46			
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.37$, df = 1 (P = 0.54); $I^2 = 0\%$						
Test for overall effect: Z = 0.40 (P = 0.69)						
1.1.8 Perinatal Death						
Ignatov et al.	0	360	0	360		Not estimable
Nunes et al.	1	3961	0	3769	10.4%	2.85 [0.12, 70.05]
The INFANT Trial	7	23263	6	23351	89.6%	1.17 [0.39, 3.48]
Subtotal (95% CI)		**27584**		**27480**	**100.0%**	**1.28 [0.46, 3.61]**
Total events	8		6			
Heterogeneity: $Tau^2 = 0.00$; $Chi^2 = 0.27$, df = 1 (P = 0.60); $I^2 = 0\%$						
Test for overall effect: Z = 0.48 (P = 0.63)						

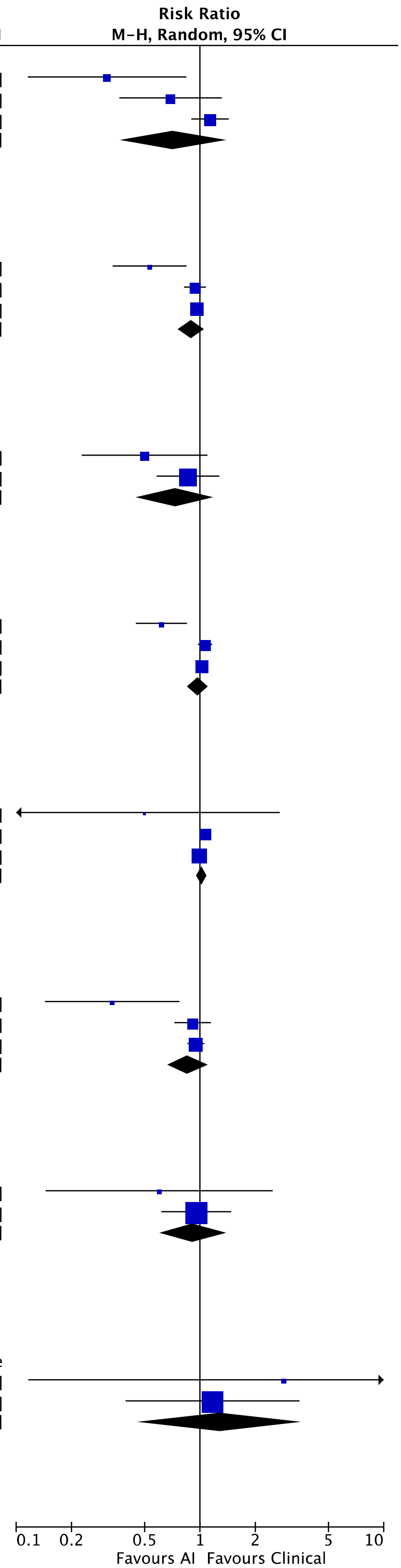


Fig. 2 Forest plot of outcomes between AI and clinical interpretation of the FHR

Table 2 Characteristics of studies included in inter-rater reliability (IRR) assessment

	Devoe et al. [19]	Bracero et al. [20]	Costa et al. [21]	Parer et al. [18]	Krupa et al. [22]	Chen et al. [23]
Year	2000	2004	2010	2010	2011	2014
Country	US	US	Portugal	US/Canada	Malaysia	Taiwan
Study design	Cohort	Cohort	Cohort	Cohort	Cohort	Cohort
#Expert observers	4	3	3	5	2	8
#Tracings assessed	50	54	50	30	90	62
AI technology	HP TraceVue	Oxford Sonicaid	Omniview-SisPorto	PeriCALM	Trium CTG Light	LabVIEW
Main Findings[a]	Cohen's $\kappa=0.25$ (0.19–0.30)	Cohen's $\kappa=0.13$ (−0.037 to 0.28)	ICC = 0.85 (0.46–0.93)	Cohen's $\kappa=0.58$ (0.48–0.68)	Cohen's $\kappa=0.68$ (0.39–0.97)	Cohen's $\kappa=0.80$ (0.67–0.94)
Other findings	The χ^2 tests applied to the levels of agreement among observers: baseline FHR, $\chi^2=3.26$ ($P=0.66$); acceleration frequency, $\chi^2=19.6$ ($P=0.001$); deceleration frequency, $\chi^2=62.5$ ($P=0.001$)	Kendall's W = 0.286	Concordant identification observed: 71% of accelerations (95% CI 69–73%), 68% of decelerations (95% CI 66–70%), and 87% of uterine contractions (95% CI 85–89%)	The clinicians agreed exactly with the majority opinion in 57% (95% confidence interval [CI], 49–64%) of the segments and were within 1 colour code in 89% (95% CI 81–96%). The average proportion of agreement was 0.83 (95% CI 0.73– 0.94)	The overall geometric mean of sensitivity and specificity was 94.8% for the training set and 81.5% for the testing set	There was good agreement for baseline variability (Cohen's $\kappa=0.68$), the numbers of early decelerations (ICC 0.78) and late decelerations (ICC 0.67), category II (kappa statistic 0.78) and moderate agreement for the number of variable decelerations (ICC 0.60), and category III (kappa statistic 0.50)
Interpretation	Fair agreement	Poor agreement	Very good agreement	Moderate agreement	Moderate agreement	Good agreement
Weighed mean Cohen's Kappa [11]	$\kappa=0.49$ [0.32–0.66]					

ICC intra-class correlation

[a]Interpretation of Cohen's κ, Kendall's W and the ICC is similar: a value of 0.0 implies no agreement, whereas a value of 1.0 implies perfect agreement

the FHR. The potential reasons why these methods have not improved outcomes are important and worth exploring.

The fundamentals of AI

First, it is fundamental to explore the basics of artificial intelligence so as to understand these results in the proper context. To create an "intelligent" system capable of processing knowledge, it has to be given means of obtaining the knowledge in the first place. Indeed, any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals is said to have some degree of AI. Human beings receive most of the information about the surrounding environment through two channels: vision and language, which includes oral and written communications. In a similar fashion, for machine learning, programming algorithms have to be provided first so that pattern recognition may then follow. In the case of EFM, an intelligent machine will learn to recognize that specific tracing patterns it has been taught—a priori, by human beings—are associated with specific neonatal outcomes.

It is then conceivable that any primitive machine, whose tasks and purpose are pre-determined and well defined, cannot outperform the knowledge of its developer, unless it is given the option to acquire new knowledge over a period of time. Efforts in studies using AI seek to determine "agreement" between observer and machines, not the "correct" interpretations. In other words, we need to ask whether the machine interprets a specific type of deceleration as the expert observer does, and if so, to what degree? In more complex AI, this is done via "reinforcement learning", where goals can be implicitly induced by rewarding some types of behavior and punishing others, or via an evolutionary system, which can induce goals by using a "fitness function" to mutate and preferentially replicate high-scoring AI systems. The most accelerated knowledge acquisition is exponential in nature, and may some point reach what is referred to as the "technological singularity"—a threshold level above which the machine becomes more capable and more intelligent than the human expert.

In the context of the trials reported in this study, these learning improvements were not meant to take place. The original purpose of AI-assisted EFM interpretation was to quantify patterns that fell in the atypical category, also known as category II. Though these are traditionally difficult to interpret, 22–39% of intrapartum tracings will see a period that meets the criteria for a category II [12]. Of these, only a small minority will develop adverse outcomes. For this reason, it is not surprising that at this preliminary stage, the AI-assisted machine did not outperform experts and did not, therefore, improve neonatal outcomes.

Electronic fetal monitoring (EFM) interpretation

Because of the known high inter-observer and intra-observer variability in the interpretation of FHR tracings, the American College of Obstetricians and Gynecologists (ACOG), the Society for Maternal–Fetal Medicine (SMFM), and the United States National Institute of Child Health and Human Development (NICHD) convened a workshop to standardize definitions and interpretation of EFM [3].

Logically, if the AI-assisted systems are meant to improve the traditional visual evaluation of EFMs, it should be first shown that the use EFM improves outcomes over other methods of fetal surveillance or over no surveillance at all. Unfortunately, this has not been consistently shown, as well conducted meta-analyses have demonstrated that EFM during labor is associated with reduced rates of neonatal seizures, but no clear differences in cerebral palsy, infant mortality or other standard measures of neonatal wellbeing, including mode of delivery, APGAR scores, NICU admissions, or academia, amongst others [7]. Likewise, large RCTs have demonstrated that the Fetal ECG ST-segment analysis used as an adjunct to conventional intrapartum electronic fetal heart-rate monitoring did not improve perinatal outcomes or decrease operative-delivery rates [13]. Although virtually all obstetrical societies advise monitoring the FHR during labor, the benefit of this intervention has not been clearly demonstrated and this position is largely based upon expert opinion and medico-legal precedent. Consequently, if the patterns programmed to be recognized by the machine aren't themselves of value to reduce neonatal morbidity, it is not surprising that adding a second evaluator with similar instructions, be it a human or a machine, does not confer any advantages in reducing adverse outcomes.

Inter-rater reliability

In this study, we showed a moderate degree of inter-rater reliability between computer and human evaluation of EFM with a weighted Cohen's Kappa of 0.49 [0.32–0.66]. In the authors' view, this is a far more important finding than the inability of AI-assisted interpretation of EFM to improve neonatal outcomes. It is well established that the IRR between human experts interpreting the FHR is poor [14, 15]. On the other hand, the rates of neonatal outcomes and complications across different obstetrical health providers show far greater homogeneity. While this is likely a consequence of the low frequency of adverse neonatal outcomes, is it in our view not out of the realm of possibilities that the interpretation of EFMs may be of little clinical utility in the general obstetrical population.

Fourier evaluation of tracings and acidosis in labor: the FETAL technique

Since appraisal of FHR tracings exhibits poor to moderate intra-observer, inter-observer and human–machine reliability, one fundamental question arises: is the key to predicting fetal wellbeing or distress embedded within the fetal heart tracing at all? And if so, does it lie beyond its traditional clinical interpretation?

There is no current way to assess correctness because there is no absolute standard for FHR interpretation. Thus, we need to rely on surrogates, such as majority opinions, and epidemiologic studies to act upon the interpretation of the real-time observations in the EFM at play. To bypass these limitations, the authors are currently developing an algorithm to overcome inter-observer variability and reduce signal-associated noise that may mask fetal compromise.

The FETAL technique, which stands for the "Fourier Evaluation of Tracings and Acidosis in Labor", is an innovative method, which applies the discrete Fourier transformation to EFM tracings to determine the spectral frequency and power distribution of the FHR. The Fourier transformation is a mathematical tool, which converts a time-dependent waveform signal, such as the fetal heart rate, into a frequency and power spectrum domain of fundamental sinusoidal waves. Operations performed in a time domain have corresponding operations in the frequency domain, which are easier to analyze and which elicit properties of the original function that are not immediately apparent on visual inspection. The mathematical details of this technique are beyond the scope of this manuscript, but, by applying the FETAL technique clinically, specific frequency distribution patterns of the FHR will be correlated with specific umbilical pH values in the newborn. Though not yet tested clinically, these frequency distributions have been theorized to provide the missing link between fetal heat rate patterns and acid–base status at birth.

Physiological basis for the application of the FETAL technique

Heart rate variability investigations started in Obstetrics, with the observation that changes in FHR variability precede changes in actual heart rate in cases of intrauterine asphyxia [16]. Indeed, the heart rate depends on the sinus node's intrinsic rate and sympathetic-parasympathetic nervous tone balance, which themselves are directly dependent on present oxygenation status [16]. Power spectral analysis and Fourier transformation of FHR variability has shown that sympathetic and parasympathetic nervous activities make frequency-specific contributions to the heart rate power spectrum, and that renin-angiotensin system activity strongly modulates the amplitude of the spectral peak located at 0.04 Hz [17]. Specifically in the non-anomalous term fetus, heart rate variability estimated by the high-frequency (HF) bands between 0.15 and 0.45 Hz reflects FHR control by the parasympathetic tone and the low-frequency (LF) bands between 0.03 and 0.15 Hz reflects the sympathetic tone [17]. Studies of heart rate fluctuation based on frequency analysis have been carried out in animal models [16–18]. However, their clinical application in humans has not been studied as of yet. Previous studies have found that the shift of autonomic balance is related to the redistribution of the power between the LF and HF bands, and that normalized power units or LF/HF are effective methods of determining the shift of autonomic balance—the mechanism presumed to be at the helm of fetal acid–base status. By studying specific low- and high-frequency patterns of Fourier-transformed domains, the FETAL Study seeks to determine whether fetal acid–base status evaluations can be made non-invasively and in real-time. Indeed, by decomposing the EFM signal in real-time into its primordial components using the Fourier transform, the FETAL technique will seek to establish a sensitive, specific and universal approach to reduce uncertainty in the interpretation of the FHR. We expect that the addition of AI-assistance in this task will enhance the sensitivity of EFM interpretation, which may manifest as an overall reduction of adverse neonatal outcomes. Table 3 showcases the implications of a successful application of the FETAL technique in a clinical setting.

Strengths and limitations

The strengths of this study are multiple and include the vast number of patients analyzed, the homogeneity of findings across studies, as well as the combination of analyses regarding both neonatal outcomes and the IRR. On the other hand, the limitations of our study are several and worth mentioning. First, while all of the studies included in the meta-analysis use AI technology, the platform used in each study was different, which may have introduced bias. Second, while the RCTs addressing the neonatal outcomes all analyzed intrapartum tracings, the studies addressing the IRR included both antepartum and intrapartum FHR tracings. Combining findings related to both clinical scenarios may impair external validity.

Table 3 Implications of a successful application of the FETAL technique

Identification of true fetal distress
Allows for adequate resuscitation or expedited delivery where appropriate
Reduces rates of NICU admission
Reduction in the rate of cesarean delivery for suspected fetal distress
Reduces future rates of repeat cesarean delivery
Reduces risk of abnormal placentation and its associated complications
Reduces maternal morbidity and mortality
Reduction in the rate of assisted vaginal delivery for suspected fetal distress
Reduces maternal morbidity
Reduces neonatal morbidity
Lower rates of intervention during labor
Enhances chances of completing a spontaneous vaginal delivery
Reduced rates of parental and provider psychological distress
Enhances therapeutic relationship between patient, family and provider
Enhances chances of completing a spontaneous vaginal delivery
Forensic analysis of tracings
Allows for education of providers
Identifies key window in labor where fetal distress is present

Conclusion

The use of AI and computer analysis for the interpretation of EFM during labor does not improve neonatal outcomes. Inter-rater reliability between experts and computer systems is moderate at best. Future studies should aim at further elucidating these findings.

Author contributions Both authors (JB, GS) accomplished all tasks equally: protocol/project development; data collection or management; data analysis; manuscript writing/editing.

Funding This study was not funded.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Verghese A, Shah NH, Harrington RA (2018) What this computer needs is a physician: humanism and artificial intelligence. JAMA 319(1):19–20
2. Hamet P, Tremblay J (2017) Artificial intelligence in medicine. Metab Clin Exp 69S:S36–S40
3. Macones GA et al (2008) The 2008 National Institute of Child Health and Human Development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. Obstet Gynecol 112(3):661–666
4. Jauniaux E, Prefumo F (2016) Fetal heart monitoring in labour: from pinard to artificial intelligence. BJOG 123(6):870
5. Lear CA et al (2016) The myths and physiology surrounding intrapartum decelerations: the critical role of the peripheral chemoreflex. J Physiol 594(17):4711–4725
6. Liston R, Sawchuck D, Young D (2018) No. 197b-Fetal health surveillance: intrapartum consensus guideline. J Obstet Gynaecol Can 40(4):e298–e322
7. Alfirevic Z et al (2017) Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. Cochrane Database Syst Rev 2:CD006066
8. Group, I.C. (2017) Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet 389(10080):1719–1729
9. Ignatov PN, Lutomski JE (2016) Quantitative cardiotocography to improve fetal assessment during labor: a preliminary randomized controlled trial. Eur J Obstet Gynecol Reprod Biol 205:91–97
10. Nunes I et al (2017) Central fetal monitoring with and without computer analysis: a randomized controlled trial. Obstet Gynecol 129(1):83–90
11. Sun S (2011) Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method 11(3–4):145–163
12. Jackson M et al (2011) Frequency of fetal heart rate categories and short-term neonatal outcome. Obstet Gynecol 118(4):803–808
13. Belfort MA et al (2015) A randomized trial of intrapartum Fetal ECG ST-segment analysis. N Engl J Med 373(7):632–641
14. Fairley S, Lawson H, Morris K (2000) Inter-observer agreement in cardiotocogram interpretation: how reliable is Ontario? J SOGC 22(5):366–373
15. Donker DK, van Geijn HP, Hasman A (1993) Interobserver variation in the assessment of fetal heart rate recordings. Eur J Obstet Gynecol 52(1):21–28
16. Parer JT et al (2006) Fetal acidemia and electronic fetal heart rate patterns: is there evidence of an association? J Matern Fetal Neonatal Med 19(5):289–294
17. Akselrod S et al (1981) Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. Science 213(4504):220–222
18. Parer JT, Hamilton EF (2010) Comparison of 5 experts and computer analysis in rule-based fetal heart rate interpretation. Am J Obstet Gynecol 203(5):451 **(e1-7)**
19. Devoe L et al (2000) A comparison of visual analyses of intrapartum fetal heart rate tracings according to the new national institute of child health and human development guidelines with computer analyses by an automated fetal heart rate monitoring system. Am J Obstet Gynecol 183(2):361–366
20. Bracero LA, Roshanfekr D, Byrne DW (2000) Analysis of antepartum fetal heart rate tracing by physician and computer. J Maternal Fetal Med 9(3):181–185
21. Costa MA et al (2010) Comparison of a computer system evaluation of intrapartum cardiotocographic events and a consensus of clinicians. J Perinat Med 38(2):191–195
22. Krupa N et al (2011) Antepartum fetal heart rate feature extraction and classification using empirical mode decomposition and support vector machine. Biomed Eng Online 10:6
23. Chen CY et al (2014) Comparison of a novel computerized analysis program and visual interpretation of cardiotocography. PLoS ONE 9(12):e112296