npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02556-y

Randomised study of human machine collaboration for cardiotocography interpretation during labour

Imane Ben M'Barek[1,2] ✉, Badr Ben M'Barek[3], Grégoire Jauvion[3], Virginia Whelehan[4], Aris Papageorghiou[4,5], Erwan Le Pennec[6] & Julien Stirnemann[2,7]

Cardiotocography (CTG) interpretation during labour is subject to high interobserver variability, limiting its performance for predicting perinatal acidaemia. This study aimed to evaluate whether computerised CTG (cCTG) assistance improves clinicians' predictive performance. In a prospective randomised multi-reader design, 211 clinicians from 23 countries were proposed to assess 100 CTG recordings (50 with pH <7.15), with or without cCTG assistance. Participants predicted the occurrence of perinatal acidaemia. cCTG assistance significantly improved overall prediction, increasing the success rate from 54.0% to 61.4% ($p < 0.01$) and sensitivity from 49.3% to 61.7% ($p < 0.01$). There was no significant difference in specificity between groups (58.7% vs 61.2%, $p = 0.14$). In discordant cases, the cCTG model was correct 67.5% of the time. Agreement and reliability between clinicians were also improved across professions, countries and levels of experience. These findings suggest that cCTG enhances the detection of perinatal acidaemia.

Cardiotocography (CTG), a non-invasive method that monitors fetal heart rate (FHR) and uterine contractions (UC), is widely used by obstetricians and midwives for intrapartum surveillance[1]. While timely detection of perinatal acidaemia is regarded as an essential component of safe intrapartum, current evidence does not conclusively demonstrate that monitoring strategies alone with CTG directly result in improved long-term neonatal outcomes. Systematic reviews have shown that although continuous CTG, compared to intermittent auscultation, may reduce the incidence of neonatal seizures, it does not significantly lower the rates of hypoxic-ischaemic encephalopathy (HIE), cerebral palsy, or perinatal death, and is associated with increased caesarean delivery rates[1,2]. The interpretation of CTG is based on visual assessment by healthcare professionals using national or international guidelines[2–5]. Thus, the process is often subjective and susceptible to inter-observer and intra-observer variability. These inconsistencies in interobserver variability can lead to clinical consequences such as overuse of intervention or missed foetal compromise, affecting clinical decision-making and patient outcomes[6–8].

To address these challenges, computerised CTG (cCTG) analysis has emerged as a promising solution[9–11]. By leveraging advanced algorithms and machine learning techniques, computerised systems aim to standardise CTG interpretation while reducing human error and variability. Previous cCTG systems were designed to predict perinatal acidaemia, as neonatal pH is a commonly used proxy of foetal wellbeing at birth[11–14]. However, cCTG is not currently used in labour ward.

In a previous work, we evaluated the accuracy of human interpretation of CTG for the prediction of perinatal acidaemia[8]. In this study, we hypothesised that AI support could enhance diagnostic accuracy and consistency in CTG interpretation. To test this hypothesis, we evaluate a cCTG assistance, named DeepCTG[15], for its effectiveness to detect perinatal acidaemia by clinicians.

Results

Descriptive analysis of the participation

A total of 211 practitioners participated in the study and collectively made 8219 assessments: 5556 with cCTG analysis and 2663 without (Table 1). The median number of assessments per participant was 19, and more than 25% of participants assessed the whole 100 cases available. The cohort included midwives ($N = 54$), obstetrician-gynaecologists ($N = 74$), and residents ($N = 83$). Midwives made the most assessments per participant (median

[1]Department of Obstetrics and Gynaecology, Assistance Publique des Hôpitaux de Paris-Beaujon, Clichy, France. [2]Université Paris Cité & Institut IMAGINE, Paris, France. [3]Genos Care, Paris, France. [4]Department of Obstetrics and Gynaecology, St George's University Hospitals NHS Foundation Trust, London, UK. [5]Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Oxford, UK. [6]CMAP, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, Palaiseau, France. [7]Department of Obstetrics and Gynaecology, Assistance Publique des Hôpitaux de Paris-Necker Enfant Malade, Paris, France. ✉e-mail: imane.benmbarek@aphp.fr

Table 1 | Description of participants

	Number of participants	Number of assessments			Number of assessments per participant Median (Q1, Q3)
		Total	With cCTG	Without cCTG	
All participants	**211**	**8219**	**5556**	**2663**	**19.0 (8.0, 100.0)**
Profession					
Midwives	54	2493	1680	813	26.5 (8.25, 100.0)
Obstetrician-Gynaecologists	74	2611	1770	841	19.0 (8.25, 66.75)
Residents	83	3115	2106	1009	17.0 (8.0, 62.5)
Country					
France	118	4420	2983	1437	17.5 (9.0, 73.0)
United Kingdom	35	1290	875	415	14.0 (7.0, 75.5)
Vietnam	29	1777	1203	574	73.0 (21.0, 101.0)
Other Europe[a]	13	384	259	125	13.0 (5.0, 48.0)
Other Asia[b]	6	142	97	45	7.0 (6.0, 16.25)
Other Africa[c]	5	148	100	48	13.0 (9.0, 20.0)
Other America[d]	5	58	39	19	6.0 (6.0, 21.0)
Years of experience					
Less than 5 years	143	5615	3797	1818	19.0 (9.0, 92.0)
More than 5 years	68	2604	1759	845	19.5 (7.0, 100.0)
Place of practice					
Academic hospital	161	6600	4458	2142	21.0 (9.0, 100.0)
Nonacademic hospital	50	1619	1098	521	10.5 (6.0, 42.25)

cCTG computerised cardiotocography.
[a]Albania, Belgium, Netherlands, Finland, Germany, Ireland, Italy, Lithuania, Spain, Switzerland, Sweden.
[b]Cambodia, Lebanon.
[c]Algeria, Gabon, Morocco, Tunisia.
[d]Brazil, Canada, Mexico.

26.5), while residents made the least (17.0). Participants were mainly from France ($N = 118$), with additional participants from the UK ($N = 35$), Vietnam ($N = 29$), and other countries ($N = 29$). Notably, participants from Vietnam achieved a median of 73 assessments each. Most participants had less than 5 years of experience ($N = 143$), and 161 worked in a university hospital.

Success rate, sensitivity, and specificity

Overall, the use of cCTG assistance significantly improved both the success rate (61.4% [60.1–62.7] vs. 54.0% [52.0–55.9], $p < 0.01$) and sensitivity (61.7% [59.8–63.7] vs. 49.3% [46.5–52.0], $p < 0.01$), with no significant difference in specificity (61.2% [59.4–63.0] vs. 58.7% [56.0–61.4], $p = 0.14$) (Supplementary Table 1). Overall, there was a disagreement between the cCTG model and human participants in 34.6% of cases (31.6% of cases when no assistance was provided and 41.0% otherwise). In these discordant cases, the AI model was correct (i.e. it correctly predicted the outcome) in 67.5% of cases (68.9% without assistance and 66.6% otherwise).

When stratified by profession, all groups benefited from cCTG assistance. Success rate increased significantly for midwives ($p < 0.01$), obstetrician-gynaecologists ($p < 0.01$), and residents ($p < 0.01$). Similarly, sensitivity improved significantly in all three groups. Across countries, significant improvements in both success rate and sensitivity were observed in France, the United Kingdom, and Vietnam (all $p < 0.05$). Regarding experience, both junior (≤5 years) and senior (>5 years) clinicians showed significant gains in success rate ($p < 0.01$ for both) and in sensitivity ($p < 0.01$ for both). Finally, in terms of practice setting, professionals from both academic and non-academic hospitals demonstrated significantly higher success rates with cCTG ($p < 0.01$ for both), along with improved sensitivity ($p < 0.01$ for both). Specificity remained unchanged across all subgroups ($p = 0.14$ and $p = 0.78$, respectively).

Success rate, sensitivity, and specificity across pH range (Fig. 1)

The success rates varied significantly across the range of pH. Figure 1 compares the performance of cCTG-assisted human assessment, unassisted human assessment and cCTG alone. Human performance with cCTG assistance was consistently higher than without assistance, particularly for pH <7.15. Notably, for severely acidemic cases (pH <7.05), there was no significant difference between the success rate achieved by cCTG alone and cCTG-assisted human assessment. However, the model still outperformed both assisted and unassisted human interpretation across all other pH ranges, achieving success rates consistently above 70%.

Agreement and reliability (Fig. 2)

Figure 2 shows the success rate per participant for predictions with and without cCTG assistance. A visual analysis suggested that cCTG smooths the success rates over participants. The standard deviation of success rates across participants decreased from 12.0% without cCTG assistance to 8.2% with cCTG assistance ($p < 0.01$).

Across all groups (profession or country), there was a significant improvement in agreement and reliability depending on whether cCTG assistance was available. For comparisons by profession, the average difference in PA and κ was 0.09 [0.01–0.17] and 0.18 [0.07–0.30], respectively. For comparisons by country, the average difference in PA and κ was 0.15 [0.05–0.25] and 0.31 [0.17–0.45], respectively. For comparisons by experience level, the average difference in PA and κ was 0.11 [0.02–0.19] and 0.23 [0.10–0.36], respectively. In all cases, the confidence intervals excluded zero, indicating a statistically significant increase in inter-rater agreement with cCTG assistance. (Supplementary Tables 2–4).

Discussion

In this study, conducted on a cohort with 8219 assessments by 211 participants, we demonstrated that cCTG assistance improved the overall success

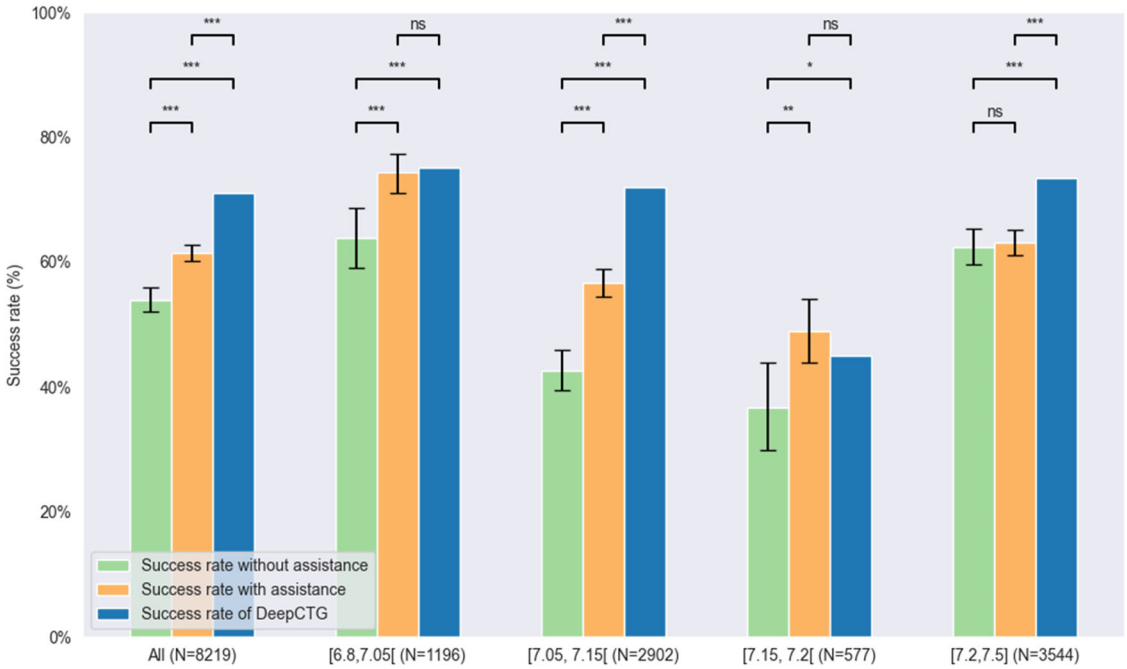

Fig. 1 | Comparison of success rates for the prediction of perinatal acidaemia across umbilical artery pH categories. Bar plots show the success rates (%) for predicting perinatal acidaemia by unassisted participants, participants assisted by computerized cardiotocography (cCTG), and the DeepCTG algorithm across ranges of umbilical artery pH. Results are shown for the entire dataset ("All") and predefined pH intervals. Error bars indicate 95% confidence intervals. Horizontal lines denote statistical comparisons between groups. ns non-significant; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

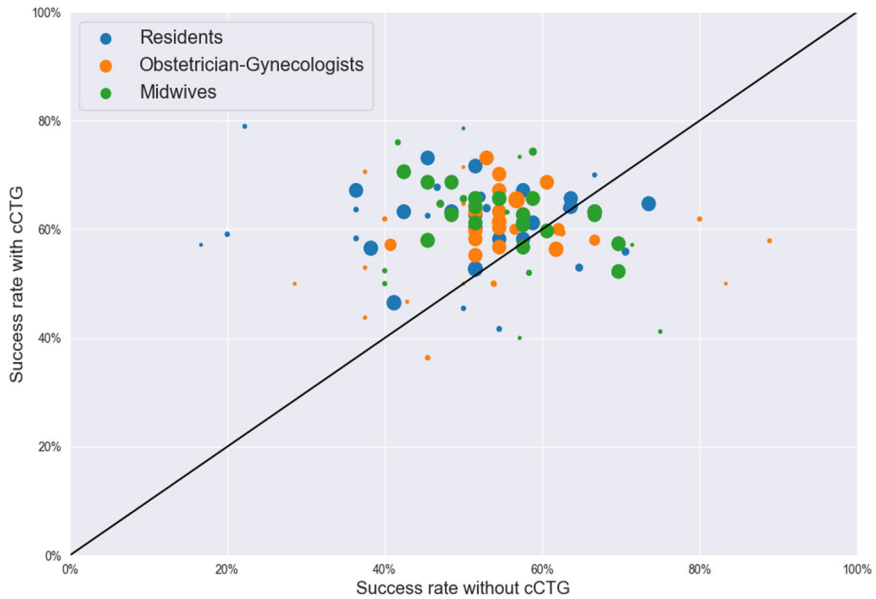

Fig. 2 | Success rate per participant without (*x*-axis) and with (*y*-axis) cCTG assistance. Each dot represents a participant, with dot size proportional to the number of cases assessed.

rate of perinatal acidaemia prediction compared to visual interpretation alone, primarily through improved sensitivity. Furthermore, cCTG assistance reduced inter-observer variability across professional backgrounds and levels of experience.

We found a significant improvement in success rate from 54.0 to 61.4% with the addition of cCTG assistance. In a similar retrospective evaluation including 204 CTG recordings[16], Costa et al. reported an increase in the success rate from 46 to 70% with cCTG analysis (Omniview-SisPorto 3.5). However, the past prospective trials have failed to demonstrate a benefit of existing cCTG solutions on perinatal morbidity or mortality. To date, the largest of such trials is the INFANT randomised controlled trial[17], which included over 47,000 women from 24 maternity units. The results showed no significant difference in adverse neonatal outcomes between groups using CTG alone or CTG combined with cCTG (adjusted relative risk 1.01, 95% CI 0.82–1.25), with a significant increase in foetal scalp pH sampling in the cCTG group (aRR 1.08 95% CI 1.01–1.16). The limitations of this study, extensively discussed elsewhere, were the study design[18,19], the performance of the model and the professionals' lack of training in using the cCTG system[17,18,20].

We found an increase in inter-observer agreement, which is also consistent with literature. On 200 STAN recordings, including 60 non-reassuring ones, Ojala et al.[21].+ found a moderate agreement (κ values

ranging from 0.47 to 0.60) among experienced obstetricians, highlighting the variability in interpretation even with standardised guidelines. In an evaluation including 30 non-reassuring CTG recordings and seven obstetricians, Vayssière et al.[22] showed that inter-observer agreement improved when practitioners used CTG combined with STAN compared with CTG alone ($\kappa = 0.67$ vs. 0.50), particularly for justified non-intervention decisions. Costa et al.[16] reported a significant improvement in inter-observer agreement when the computer analysis was available, with an intra-class correlation coefficient (ICC) rising from 0.43 to 0.70 in the cCTG-assisted group.

In cases where there was a disagreement between the participant and the model, the model's prediction was correct in 67.5% of cases, suggesting that the algorithm may provide valuable corrective input, particularly when human interpretation is uncertain or incorrect.

In our study, the greatest additional value of cCTG assistance was found among less experienced participants, allowing them to reach a performance close to that of experienced clinicians[6,23,24]. This standardisation of CTG interpretation is also shown by the improvement in inter-observer agreement across participants from different professions and countries. Our results also suggest that cCTG assistance may help clinicians better detect perinatal acidaemia by increasing sensitivity, without leading to an increase in unnecessary interventions, as specificity was maintained. While the absolute improvement in success rate from 54.0% to 61.4% may appear modest, it remains clinically meaningful, particularly in high-risk populations, where each gain in predictive accuracy can help prevent a potentially serious and avoidable adverse outcome.

Our findings contribute to the growing field of human–machine collaboration, aiming at combining human expertise with AI capabilities to enhance clinical decision-making. Several studies have shown that AI systems alone often outperform clinicians, even when used in a collaborative setting, leading to suboptimal decisions[25,26]. This suggests that physicians tend to undervalue AI recommendations or to stick to their initial impressions. Prinster et al.[27] evaluated how using an AI model along with a confidence score and detailed explanations helped physicians interpret chest radiographs and concluded to an improved diagnostic accuracy and an increased trust in the AI model. These results suggest that cCTG must be evaluated by its ability to improve human practitioners[25,26], and to be implemented effectively in clinical practice, it should consist of a prediction model along with a measure of confidence and interpretable features.

Several modelling strategies can be considered when developing a diagnostic assistant such as ours: the model used in our study displayed convincing performance while being relatively simple (based on a logistic regression) and providing explainable features along with a prediction, which could be a significant factor for adoption by clinicians. More advanced modelling approaches (e.g. deep learning models) could perhaps reach higher performance, but with the loss of part of the explainable clinical features of the prediction, possibly limiting the adoption by clinicians.

A major strength of our study lies in its large sample size and the use of a validated, publicly available database. Our setup enabled flexible participation and facilitated large-scale data collection across a broad spectrum of professional backgrounds and geographies. Moreover, the cCTG system was based on a validated model[15] integrated into a user-friendly web-based tool. The higher number of assessments performed by participants from certain countries (such as Vietnam) highlights the strong engagement with the tool. Beyond AI-assisted prediction, it provides immediate feedback and interpretative guidance, which may support its use as a complementary educational resource for CTG interpretation.

Several limitations must be acknowledged. First, our analysis focused solely on acidaemia as the outcome, which does not capture the full spectrum of perinatal risks associated with labour and delivery, and serves only as a proxy for neonatal and long-term morbidity[13,28,29]. We chose a threshold of pH = 7.15, to include moderate levels of acidaemia (pH between 7.05 and 7.20), with the aim of identifying situations where timely intervention could prevent progression to more severe acidaemia. Large cohort studies suggest that even moderate acidaemia may already be associated with increased neonatal morbidity and mortality, underscoring the potential clinical relevance of earlier detection[12,28,30]. Nonetheless, subgroup analyses using lower thresholds showed that the improvement observed with cCTG is consistent across the range of observed pH. Second, the assessment setup used in our study differs from typical practices in the labour ward. One major difference is the rate of adverse outcomes that has been artificially increased to enable robust statistical analysis. Another difference comes from the exclusion of caesarean deliveries in the study. Further prospective studies should evaluate the implementation of such assistance in clinical practice. Third, even if our results are overall consistent with the literature, the comparison with similar studies was challenging[31]. A key reason lies in the variability of cCTG systems: some focus on foetal ECG analysis[21,22,32], others evaluate specific signal features[33–36], and some aim to classify the risk of acidaemia[32,37] or predict broader clinical outcomes[16,38].

To conclude, this study suggests that cCTG could assist clinicians in improving the detection of perinatal acidaemia and reducing variability in CTG interpretation, without an increase in false positives that could increase the risk of unnecessary interventions.

Methods

Participants

Participants were obstetricians and midwives, all clinically involved in deliveries, recruited in 23 countries (France, UK, Vietnam, Albania, Belgium, Netherlands, Finland, Germany, Ireland, Italy, Lithuania, Spain, Switzerland, Sweden, Cambodia, Lebanon, Algeria, Gabon, Morocco, Tunisia, Brazil, Canada, Mexico) from January 2024 to March 2025. The age, profession (midwives, obstetrician-gynaecologists and residents), years of experience, type of practice (university or general hospital), and country were recorded for each participant.

Study protocol

We conducted a prospective randomised multi-reader study based on a retrospective database. We developed a website (www.fhr-annotator.com) designed as a platform for online testing of our tool. The website was widely disseminated via email, both through professional medical associations and by directly contacting the heads of obstetrics and gynaecology departments. For every case, the last 45 min before delivery of the CTG signals (FHR and UC) were given along with some clinical information including sex, gestational age and birth weight. Because CTG paper speed varies according to national guidelines and standard clinical practice, participants were allowed to use either 1 cm/min or 3 cm/min depending on the format routinely used in their country. Based on the CTG and clinical information, participants were asked to predict the neonatal umbilical artery pH (either normal or <7.15). This prediction is then compared to the true outcome based on the umbilical artery pH at birth, and the participants were notified whether the prediction was correct or not after each case. Participants could log in and out as needed, but could not go back to correct previous predictions. All predictions were recorded.

The clinical cases presented to the participants were all taken from the open-source CTU-UHB dataset[39]. This dataset contains 552 deliveries >37 weeks, for which CTG signals (FHR and UC) and umbilical cord blood pH are available. This database comprises 105 cases with a pH <7.15.

We randomly extracted 100 cases from this dataset: 50 cases with pH below 7.15 (indicating moderate or severe perinatal acidaemia) and 50 cases with pH above 7.15 (indicating a normal outcome). We chose a pH threshold of 7.15 to define perinatal acidaemia, consistent with previous research showing its relationship with an increased risk of HIE[12,40]. Deliveries by caesarean were excluded to ensure that no relevant information was missing at the end of the recording, before delivery (i.e. the time between the end of CTG and delivery is unknown in case of a caesarean).

The 100 cases were split into 10 batches of 10 cases each, with every batch containing 5 cases with a normal outcome and 5 cases with perinatal acidaemia (pH < 7.15). The batches were presented in the same order to all participants, but the order of cases within each batch was randomly shuffled. Furthermore, AI assistance was randomly allocated to 2/3 of cases,

Fig. 3 | Example of a clinical case displayed to participants with computerized cardiotocography (cCTG) analysis. Screenshot of the interface used by participants during the annotation task. Clinical characteristics are displayed at the top of the screen together with the outcome options. The model prediction panel indicates the predicted risk of neonatal acidaemia and the main contributing features, as predicted by DeepCTG. The lower panel shows the CTG tracing with extracted CTG variables, including deceleration areas (red), acceleration areas (green), and the model-derived fetal heart rate baseline (red dashed line).

differently so for each participant. This approach ensured both outcomes are balanced within a batch and a greater consistency in the case sets reviewed across participants, compared to fully random allocation. In two-thirds of cases, the output of the cCTG analysis was given to the participant. This unbalanced distribution was deliberately chosen to enhance the learning curve with the decision support system. The remaining third was assessed without cCTG analysis to evaluate the impact of cCTG analysis on the accuracy of the prediction. cCTG analysis is composed of four levels of information: (i) a visual representation of the FHR baseline, and areas of acceleration and deceleration on the CTG reading, (ii) a prediction of the probability of perinatal acidaemia defined by pH <7.15 (from 0 to 100%), (iii) categories derived from this risk: low (<30%), moderate (30–80%) and high (>80%), and (iv) a decomposition of the risk on a few explanatory variables (FHR baseline, and areas of acceleration and deceleration). This analysis was performed with DeepCTG®1.0, a logistic regression that predicts a risk of perinatal acidaemia based on four variables extracted from the last 30 min of CTG signals before delivery: the minimum and maximum values of FHR baseline, and the total area of accelerations and decelerations (in cm^2). Participants were blinded to the correctness of the cCTG. The AI system was previously validated on the same CTU-UHB dataset, with performance metrics reported elsewhere[15], along with the methodology used to decompose the risk across the explanatory variables. Figure 3 shows how cases are displayed to the participants with cCTG analysis.

Data analysis

The accuracy of predictions of perinatal acidaemia was measured through three metrics: (i) success rate (the proportion of correct predictions—either correctly predicting perinatal acidaemia or its absence); (ii) sensitivity (the proportion of predictions that were correctly classified as perinatal acidaemia) and (iii) specificity (the proportion of predictions that were correctly classified as normal outcome). These metrics were computed separately over assessments with and without cCTG assistance, in the whole sample of participants/raters as well as by profession, country, years of experience and place of practice. We also computed the metrics using cCTG alone, considering that the prediction of the cCTG model was perinatal acidaemia when the predicted probability was higher than 50%, and normal otherwise. Success rates were computed over the range of neonatal pH, broken down as follows: <7.05; 7.06–7.14; 7.15–7.19; >7.20.

All metrics were accompanied by their 95% confidence intervals.

Following the GRRAS guidelines[41], we used the proportion of agreement (PA) to assess the level of agreement between participants based on their profession, years of experience and country. To further evaluate the reliability of this agreement beyond chance, we calculated Cohen's Kappa coefficient (κ)[3,41]. The assessments were separated in groups based on profession, years of experience and country. Within each group and for each of the 100 cases, we determined the most frequent prediction (perinatal acidaemia or normal outcome). Agreement and reliability between two

groups (e.g. midwives and residents) were evaluated by comparing the most frequent prediction from the two groups (with their confidence intervals) over the 100 cases. The results with and without cCTG assistance were compared.

We also derived an overall measure for agreement and reliability by averaging the agreements and reliabilities between all pair of groups. These overall measures were computed separately for groups based on either profession, years of experience or country. We evaluated the difference in those overall measures depending on whether cCTG assistance was used, with a confidence interval. The difference was considered significant if the 95% confidence interval did not include zero[42].

Interpretation of PA followed commonly accepted thresholds: values <0.40 were considered poor agreement, 0.40–0.59 fair, 0.60–0.74 good, and ≥0.75 excellent agreement[43]. Moreover, if the lower boundary of the 95% CI for PA fell below 0.50, the agreement level was also regarded as non-significant[44]. For Cohen's Kappa, values ≤ 0 indicated no agreement, 0.01–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1.00 almost perfect agreement[45].

To assess the impact of cCTG assistance on inter-individual variability, we computed the standard deviations of success rate over the whole sample of participants, with and without cCTG assistance. The standard deviations with and without cCTG assistance were compared using a two-sided F-test.

We used Python 3.12.8 for the analysis along with libraries 'scikit-learn' and 'statsmodels' for statistical analysis, 'pandas' for data processing and 'matplotlib' for visualisation.

This work obtained the ethical approval of Robert Debré ethical committee, Assistance Publique-Hôpitaux de Paris (CEER-RD 2023-653).

All participants were informed on the web application before the start of the survey and gave their consent to the study before starting their evaluation sessions.

Data availability

The datasets generated during the current study are not publicly available as the consent obtained did not include public data sharing but are available from the corresponding author on reasonable request.

Received: 24 June 2025; Accepted: 6 March 2026;
Published online: 19 March 2026

References

1. Alfirevic, Z., Gyte, G. M., Cuthbert, A. & Devane D. Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst. Rev.* 2017. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6464257/.
2. Ayres-de-Campos, D., Spong, C. Y. & Chandraharan, E. FIGO consensus guidelines on intrapartum fetal monitoring: cardiotocography. *Int. J. Gynecol. Obstet.* **131**(1), 13–24 (2015).
3. Santo, S. et al. Agreement and accuracy using the FIGO, ACOG and NICE cardiotocography interpretation guidelines. *Acta Obstet. Gynecol. Scand.* **96**(2), 166–175 (2017).
4. Carbonne, B. et al. Classification CNGOF du rythme cardiaque fœtal : obstétriciens et sages-femmes au tableau. *J. Gynécol.Obstétr. Biol. Reprod.* **42**, 509–510 (2013).
5. Chandraharan, E. Introduction of the physiological CTG interpretation & hypoxia in labour (HIL) Tool, and its incorporation into a software programme: impact on perinatal outcomes. *Glob. J. Reprod. Med*. **8**, 8 (2021).
6. Blackwell, S. C., Grobman, W. A., Antoniewicz, L., Hutchinson, M. & Gyamfi Bannerman, C. Interobserver and intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system. *Am. J. Obstet. Gynecol.* **205**(4), 378.e1–5 (2011).
7. Hruban, L. et al. Agreement on intrapartum cardiotocogram recordings between expert obstetricians. *J. Eval. Clin. Pr.* **21**(4), 694–702 (2015).
8. Ben M'Barek, I. et al. Large-scale analysis of interobserver agreement and reliability in cardiotocography interpretation during labor using an online tool. *BMC Pregnancy Childbirth* **24**(1), 136 (2024).
9. Mendis, L., Palaniswami, M., Brownfoot, F. & Keenan, E. Computerised cardiotocography analysis for the automated detection of fetal compromise during labour: a review. *Bioengineering* **10**(9), 1007 (2023).
10. Jones, G. D., Cooke, W. R., Vatish, M. & Redman, C. W. G. Computerized analysis of antepartum cardiotocography: a review. *Matern. Fetal Med.* **4**(2), 130 (2022).
11. Ben M'Barek, I., Jauvion, G. & Ceccaldi, P. F. Computerized cardiotocography analysis during labor – A state-of-the-art review. Acta Obstetricia et Gynecologica Scandinavica [Internet]. [cited 2022 Dec 21];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/aogs.14498.
12. Olofsson, P. Umbilical cord pH, blood gases, and lactate at birth: normal values, interpretation, and clinical utility. *Am. J. Obstet. Gynecol.* **228**(5), S1222–S1240 (2023).
13. Myrhaug, H. T. et al. Umbilical cord blood acid-base analysis at birth and long-term neurodevelopmental outcomes in children: a systematic review and meta-analysis. *BJOG* **130**(10), 1156–1166 (2023).
14. Petrozziello, A., Redman, C., Papageorghiou, A. T., Jordanov, I. & Georgieva, A. Multimodal convolutional neural networks to detect fetal compromise during labor and delivery. *IEEE Access* **5**, 1 (2019).
15. Ben M'Barek, I. et al. DeepCTG® 1.0: an interpretable model to detect fetal hypoxia from cardiotocography data during labor and delivery. Frontiers in Pediatrics [Internet]. 2023 [cited 2023 Jun 21];11. Available from: https://www.frontiersin.org/articles/10.3389/fped.2023.1190441.
16. Costa, A., Santos, C., Ayres-de-Campos, D., Costa, C. & Bernardes, J. Access to computerised analysis of intrapartum cardiotocographs improves clinicians' prediction of newborn umbilical artery blood pH. *BJOG* **117**(10), 1288–1293 (2010).
17. Brocklehurst, P. et al. The INFANT trial. *Lancet* **390**(10089), 28 (2017).
18. Belfort, M. A. & Clark, S. L. Computerised cardiotocography—study design hampers findings. *Lancet* **389**(10080), 1674–1676 (2017).
19. Keith, R. The INFANT study—a flawed design foreseen. *Lancet* **389**(10080), 1697–1698 (2017).
20. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**(1), 114–126 (2015).
21. Ojala, K., Mäkikallio, K., Haapsamo, M., Ijäs, H. & Tekay, A. Interobserver agreement in the assessment of intrapartum automated fetal electrocardiography in singleton pregnancies. *Acta Obstet. Gynecol. Scand.* **87**(5), 536–540 (2008).
22. Vayssière, C. et al. Inter-observer agreement in clinical decision-making for abnormal cardiotocogram (CTG) during labour: a comparison between CTG and CTG plus STAN. *BJOG Int. J. Obstet. Gynaecol.* **116**(8), 1081–1088 (2009).
23. Blix, E., Sviggum, O., Koss, K. S. & Øian, P. 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), 1–5 (2003).
24. Epstein, A. J. et al. Interobserver reliability of fetal heart rate pattern interpretation using NICHD definitions. *Am. J. Perinatol*. **16**, 463–468 (2012).
25. Goh, E. et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Netw. Open* **7**(10), e2440969 (2024).
26. Vaccaro, M., Almaatouq, A. & Malone, T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* **8**(12), 2293–2303 (2024).
27. Prinster, D. et al. Care to explain? AI explanation types differentially impact chest radiograph diagnostic performance and physician trust in AI. *Radiology* **313**(2), e233261 (2024).
28. Andersson, C. B. et al. Umbilical cord pH levels and neonatal morbidity and mortality. *JAMA Netw. Open* **7**(8), e2427604 (2024).

29. Baalbaki, S. H. et al. Predicting long-term neurodevelopmental outcomes in very preterm neonates by umbilical cord gas parameters. *Am. J. Obstet. Gynecol. MFM* **3**(1), 100248 (2021).
30. Bailey, E. J. et al. Mild neonatal acidemia is associated with neonatal morbidity at term. *Am. J. Perinatol.* **38**(S 01), e155–e161 (2021).
31. Hernandez Engelhart, C. et al. Reliability and agreement in intrapartum fetal heart rate monitoring interpretation: a systematic review. *Acta Obstet. Gynecol. Scand.* **102**(8), 970–985 (2023).
32. Westerhuis, M. E. M. H. et al. Inter- and intra-observer agreement of intrapartum ST analysis of the fetal electrocardiogram in women monitored by STAN. *BJOG* **116**(4), 545–551 (2009).
33. Gagnon, R., Campbell, M. K. & Hunse, C. A comparison between visual and computer analysis of antepartum fetal heart rate tracings. *Am. J. Obstet. Gynecol.* **168**(3 Pt 1), 842–847 (1993).
34. Devoe, L. et al. 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 (2000).
35. Chen, C. Y., Yu, C., Chang, C. C. & Lin, C. W. Comparison of a novel computerized analysis program and visual interpretation of cardiotocography. *PLOS ONE* **9**, e112296 (2014).
36. Magawa, S. et al. Intrapartum cardiotocogram monitoring between obstetricians and computer analysis. *J. Matern. Fetal Neonatal Med.* **34**(5), 787–793 (2021).
37. Schiermeier, S., Westhof, G., Leven, A., Hatzmann, H. & Reinhard, J. Intra- and interobserver variability of intrapartum cardiotocography: a multicenter study comparing the FIGO classification with computer analysis software. *Gynecol. Obstet. Investig.* **72**(3), 169–173 (2011).
38. Keith, R. D. et al. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *Br. J. Obstet. Gynaecol.* **102**(9), 688–700 (1995). Sep.
39. Chudáček, V. et al. Open access intrapartum CTG database. *BMC Pregnancy Childbirth* **14**, 16 (2014).
40. DuPont, T. L. et al. Short-term outcomes of newborns with perinatal acidemia who are not eligible for systemic hypothermia therapy. *J. Pediatr.* **162**(1), 35–41 (2013).
41. Kottner, J. et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J. Clin. Epidemiol.* **64**(1), 96–106 (2011).
42. Cumming, G. & Finch, S. Inference by eye: confidence intervals and how to read pictures of data. *Am. Psychol.* **60**(2), 170–180 (2005).
43. Altman, D. *Practical Statistics for Medical Research [Internet]* 404–408 (Chapman and Hall, 1991). Available from: https://www.routledge.com/Practical-Statistics-for-Medical-Research/Altman/p/book/9780412276309.
44. Grant, J. M. The fetal heart rate trace is normal, isn't it? Observer agreement of categorical assessments. *Lancet* **337**(8735), 215–218 (1991).
45. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**(1), 159–174 (1977).

Acknowledgements
The authors would like to thank all the participants of the study across the world, and the clinical research unit, especially Sabrina Verchere, for their regulatory support. No funding was received for this research.

Author contributions
This study was designed by I.B., G.J., J.S. and E.L. The website was built by B.B., the algorithm to determine the baseline was applied by GJ. Statistical analysis was performed by I.B., B.B. and G.J. The first draft of this manuscript was written by I.B. J.S, E.L., V.W. and A.P. revised it critically and all authors have been involved in the final manuscript through constructive criticism and validated the manuscript.

Competing interests
G.J. and B.B. are shareholders of Genos Care. J.S. and E.L. are cofounders of Sonio. All other authors declare no competing interests. The present study received no financial support from these companies, and no funding or grant was obtained for this work. The other authors do not have a competing interest.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02556-y.

Correspondence and requests for materials should be addressed to Imane Ben M'Barek.