Artificial intelligence-based prediction of second stage duration in labor: a multicenter retrospective cohort analysis

Xiaoqing Huang,[a,b,f] Xiaodan Di,[c,f] Suiwen Lin,[a,b] Minrong Yao,[d] Suijin Zheng,[e] Shuyi Liu,[a,b] Wayan Lau,[a,b] Zhixin Ye,[a,b] Zilian Wang,[a,b] and Bin Liu[a,b,*]

[a]Department of Obstetrics and Gynecology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[b]Guangdong Provincial Clinical Research Center for Obstetrical and Gynecological Diseases, Guangzhou, China
[c]Department of Obstetrics and Gynecology, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, China
[d]Department of Obstetrics and Gynecology, Sanming First Hospital, Sanming, China
[e]Department of Obstetrics and Gynecology, Houjie Hospital of Dongguan, Dongguan, China

eClinicalMedicine
2025;80: 103072

Published Online xxx
https://doi.org/10.1016/j.eclinm.2025.103072

Summary

Background Duration of second stage of labor is crucial for fetal delivery, but the optimal length of this stage remains controversial. While extending the duration of second stage can reduce primary cesarean delivery rates, it may increase maternal and neonatal morbidities as the duration progresses. We aimed to develop a personalized machine learning (ML) model to predict the possible second-stage duration.

Methods This multicenter, retrospective study was conducted at four tertiary hospitals in China from September 2013 to October 2022. Data from three hospitals in Guangdong Province was selected as derivation set, and a geographically independent dataset from Fujian Province as the external validation set. Singleton vaginal deliveries with term live birth in a cephalic position were included. The primary outcome was the duration of the second stage of labor. Since durations beyond 3 h were rare, we developed binary classification models with thresholds at 1 h and 2 h. After the optimal features selected by recursive feature elimination (RFE) method, four ML algorithms were employed to build the models. The best model would be selected with the predictive performance and interpreted with Shapley Additive exPlanations method. The study is registered in Clinical Trial (ChiCTR2400085338).

Findings Electronic medical records of 79,381 vaginal deliveries were obtained, and 63,401 deliveries meeting the inclusion criteria were included in the final analysis. Eight risk features were selected through the RFE process. Gradient boosting machine implemented by decision tree models achieved the best performance, yielding areas under the curve for 1-h and 2-h models of 0.808 (95% confidence interval [CI] 0.797–0.819) and 0.824 (95% CI 0.804–0.843) in the testing set, and 0.862 (95% CI 0.854–0.870) and 0.859 (95% CI 0.843–0.875) in the external validation set, respectively.

Interpretation An explainable and reliable ML model was developed to predict the probable second-stage duration, which could assist in individualized labor management. Factors such as first-stage duration and maternal age are potential predictors for the second stage.

Funding National Natural Science Foundation of China (No.82371689, N0.81771602), and National Key Research and Development Program of China (No.2021YFC2700703).

Keywords: Duration of second stage of labor; Machine learning; Personalized medicine; Obstetric prediction; Vaginal delivery

Introduction

The duration of labor is closely associated with maternal and neonatal outcomes.[1] Over the past half-century, obstetricians and midwives applied the Friedman labor pattern[2] to manage labor duration. In 2010, Zhang et al. introduced a contemporary labor pattern, which allows a

*Corresponding author. Department of Obstetrics and Gynecology, The First Affiliated Hospital of Sun Yat-sen University, 58 Zhongshan Road II, Guangzhou, 510080, China.
E-mail address: Liubn@mail.sysu.edu.cn (B. Liu).
[f]These authors contributed equally to this work.

Research in context

Evidence before this study

We searched PubMed with the terms "second stage of labor" AND ("machine learning" OR "artificial intelligence" OR "prediction") published from database inception up to October 1, 2024, with no language restrictions. There have been few studies on machine learning (ML) models related to second stage of labor, and they all aimed to predict the possibility of vaginal delivery during labor.

Added value of this study

To our knowledge, this is the first report of an artificial intelligence predictive model for the duration of second stage of labor with good performance. We provide an alternative tool to quantify the duration of second stage, aiding in individual assessment when considering the trade-offs between maternal and neonatal outcomes. We have identified and quantified clinical features that contribute to predicting the second-stage duration, including the duration of first stage and maternal age.

Implications of all the available evidence

The second stage of labor is a challenging time for pregnant women, fetuses, and care providers. As the duration of second stage increases, opportunities for vaginal delivery decrease while maternal and fetal morbidity increases. Although guidelines provide recommendations on the upper limit of the second-stage duration, there is no predictive method for clinical management. Models that use information available before second stage to predict its possible duration would be valuable. This study demonstrated that using ML models to personally predict the duration of labor is feasible. Our explainable and reliable ML model, and the online calculator based on it, can serve as valuable tools for predicting probable second-stage duration, which would significantly aid in individualized labor management by balancing the risks to maternal and neonatal outcomes and improving clinical decision-making.

longer second-stage labor duration to prevent primary cesarean deliveries.[3] However, longer second-stage labor duration increased maternal and fetal morbidities, including postpartum hemorrhage (PPH), chorioamnionitis, third- or fourth-degree perineal lacerations, and neonatal birth injuries.[1,4–8]

To guide the management of labor, both global and regional clinical practice guidelines established the upper limit of second-stage labor duration,[9–16] with some variations.[1,17–19] Moreover, individualized management of the second stage of labor can improve maternal and neonatal outcomes.[12,15] In clinical practice, obstetricians and midwives assess labor duration according to clinical characteristics, but less quantified prediction tool was applied. Recently, machine learning (ML) models have been built to predict delivery mode[20–23] and obstetric complications[24–27] in labor. ML algorithms have advantages over traditional statistics in handling high-dimensional large data,[21,25,28] but no ML model has been reported to predict the duration of second stage.

The purpose of the present study is to develop ML models to achieve the personal prediction for probable duration of second stage. We hypothesized that applying ML approach to real-world data will provide an accurate prediction of second-stage labor duration.

Methods

Ethics statement

Institutional ethical review committee approval by The First Affiliated Hospital of Sun Yat-sen University was obtained for this study (approval number: [2022]451). Given the retrospective design of the study, the requirement for informed consent was waived. The study is registered in Clinical Trial (ChiCTR2400085338). This study followed the TRIPOD guidelines for predictive model development and validation.[29]

Study design and data collection

This was a multicenter retrospective study that based on four tertiary hospitals in China from September, 2013 to October, 2022. The derivation set was consisted of three hospitals from Guangdong Province including The First Affiliated Hospital of Sun Yat-sen University (FAH), Guangzhou Women and Children's Medical Center, and Houjie Hospital of Dongguan, which was then randomly split into 80% training set and 20% testing set (internal validation set) for model development. For external model validation, a geographically independent dataset was collected from Sanming First Hospital in Fujian Province. The management of second stage in each of the institutions was strictly under the expert consensus and practice guidelines in China.[3,9,30] Generally, when labor progress is arrested, oxytocin augmentation is recommended for adequate contractions if maternal and neonatal status remains reassuring. More details of participating hospitals were listed in Supplementary Table S1.

Live singleton vaginal deliveries with cephalic position were recruited in the study. Operative vaginal deliveries, multiple pregnancy, preterm birth, antepartum or intrapartum birth death, and prior uterine surgery were excluded from the analysis. Cases lack of completed records on delivery duration were also excluded.

Clinical data collected from electronic medical records included maternal demographic features (e.g.,

age, parity, maternal body mass index [BMI], gestational age), prenatal complications (e.g., gestational diabetes mellitus, hypertensive disorders), details of the labor process (e.g., duration of labor, induction of labor, epidural analgesia, fetal head position), and maternal and neonatal outcomes (e.g., PPH, Apgar scores).

Outcome

The primary outcome of the study was the duration of the second stage of labor, defined as the period from complete cervical dilatation to the delivery of the fetus. We aimed to develop models to predict whether the second stage would exceed 1 h or 2 h, investigated as binary outcomes. Due to the insufficient number of cases with a second stage duration exceeding 3 h (186 out of 63,401, 0.29%), we were unable to develop a statistically robust model for this threshold.

Model development

During data preprocessing, the StandardScaler from the Scikit-learn package was applied for feature scaling. Due to the imbalanced data, undersampling was performed on the training set. For variables with less than 10% missing values, KNN imputation was used to supplement missing values for continuous variables,[31] and the mode was imputed for categorical variables.[32] The characteristics before and after imputation showed no significant difference both in derivation and validation set (Supplemental Table S2–S4). The complete steps of feature processing are presented in the Supplemental Figure S1.

Relevant features were collected based on experience and literature reviews.[9–12,33] To improve the models' accuracy, candidate predictors that were statistically significant in univariate tests were chosen and then screened for multicollinearity (variance inflation factor [VIF] ≥10). Furthermore, the recursive feature elimination (RFE) method,[34] a popular feature selection algorithm, was used to select the best features. RFE recursively removed the least important features, considering progressively smaller sets of features until the optimal number of features was reached, aiming to achieve similar performance to the model with the full feature set.

Four supervised ML methods were selected to create classification models: gradient boosting machine implemented by decision tree (GB), random forest (RF), K-nearest neighbor (KNN), and Gaussian Naive Bayes (GNB). We employed grid search with 10-fold cross-validation (CV) to determine the best parameters for each model.[35] The discrimination of models was measured by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, accuracy, sensitivity, specificity and F1 score.[28] The calibration of models was measured by calibration curve.

To avoid the ML models being a "black box", SHapley Additive exPlanations (SHAP) was implemented to explain the output of the models by calculating the contribution of each feature.[36]

To further explore the influence of other clinical features on the second stage, we conducted additional analysis in the delivery cohort from the FAH cohort, as it provided more detailed clinical data. Four models were sequentially developed by adding interested features. Model 1: included features used in the whole cohort; Model 2: Model 1 + maternal weight; Model 3: Model 2 + oxytocin augmentation; Model 4: Model 3 + caput succedaneum. All the adding features were subjected to RFE model to minimize potential overfitting brought by high dimensionality. As GB model showed better performance in our pilot study, we chose GB algorithm to develop FAH models according to the methods described upon.

Statistical analysis

Statistical analysis was performed with Python 3.8.8 and R 4.3.2. Scikit-learn 1.0.1 package was used to develop the models. Kolmogorov–Smirnov test was used to assess the data distribution. Continuous features were compared using the independent sample t test if normal distributed, else the Mann–Whitney U test. Dichotomous features were compared using the Chi-square test or Fisher's exact test. DeLong's method was employed to compare the AUCs of models and calculate confidence intervals (CI). A P value < 0.05 in two-tailed tests was considered statistically significant.

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A total of 79,381 vaginal deliveries were collected from four medical centers during the study period. As presented in Fig. 1, 15,980 cases were excluded due to operative vaginal deliveries, multiple pregnancy, gestational age less than 37 weeks, stillbirth, prior uterine surgery, non-cephalic presentation, or insufficient records of the second stage. Finally, 63,401 cases were analyzed in the delivery cohort, with 44,391 cases in the derivation set from Guangdong Province and 19,010 cases in the validation set from Fujian Province.

Baseline characteristics of the cohort are presented in Table 1. The mean maternal age of this delivery cohort was 29.50 years. Among the <1 h, 1–2 h, and ≥2 h groups, the rates of advanced maternal age (≥35 years), multiparity, and low birthweight infants gradually decreased, while the incidence of assisted reproductive technology, premature rupture of membranes (PROM), epidural analgesia, and artificial rupture of membranes increased. The duration of first stage, gestational weeks, and birthweight were progressively

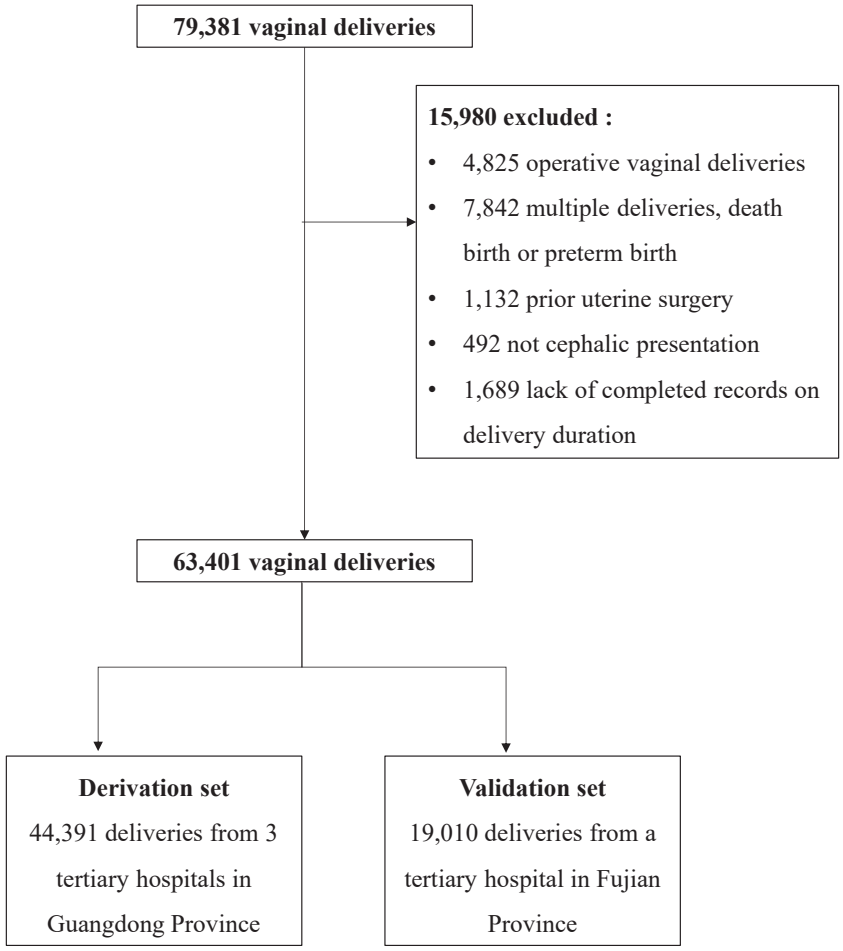

Fig. 1: **The flow chart of delivery cohort**.

higher among the three groups. The baseline characteristics of participants from each hospital were shown in Supplemental Table S5. Because of the geographically difference, participants in derivation set were slightly younger, more multipara, with less epidural analgesia and shorter second stage of labor.

The distribution of the duration of second stage was demonstrated in Fig. 2. Most women (97.70%) delivered within 2 h of second stage, and the majority of multiparous women (97.72%) had a second stage within 1 h. The median duration of the second stage of labor was 35 min (95th percentile [P95th]: 114 min) for nulliparous women and 14 min (P95th: 44 min) for multiparous women. To further understand the current labor data, the duration of second stage was stratified by parity and epidural analgesia, as shown in Supplemental Table S6, and was comparable to several prior studies.

Clinical features used in the prediction model were chosen via the selection process (Supplemental Figure S2). Twenty-nine features were collected as candidate predictors. Among them, nine were excluded due to missing data more than 10%, and four were excluded because no statistical difference was found. In the remaining sixteen features, two were excluded for multicollinearity. Consequently, eight features were selected by RFE to build the models, including maternal age, parity, gestational weeks, duration of first stage of labor, epidural analgesia, PROM, non-occiput anterior (non-OA) fetal position, and artificial rupture of membranes. As shown in Table 2, these features had higher incidences in the longer duration groups (≥1 h and ≥2 h) with statistically significant differences.

The performance measures of the ML models are presented in Table 3, after hyper parameters tuning using grid search (Supplemental Table S7). Among 1 h models, GB model demonstrated the highest sensitivity (0.847 for testing set, 0.937 for validation set), negative predictive value (NPV; 0.957 for testing set, 0.992 for validation set), F1 score (0.455 for testing set, 0.281 for validation set) and AUC (0.808 [95% CI 0.797–0.819] for testing set, 0.862 [95% CI 0.854–0.870] for validation set, Fig. 3A and B). Among 2 h models, the AUC of the

	Total	Derivation set (N = 44,391)				Validation set (N = 19,010)			
		<1 h group	1–2 h group	≥2 h group	P	<1 h group	1–2 h group	≥2 h group	P
N	63,401	37,343 (84.12)	5864 (13.21)	1184 (2.67)		17,603 (92.60)	1134 (5.96)	273 (1.44)	
Maternal age (year)	29.50 ± 4.46	30.05 ± 4.38	29.34 ± 3.81	29.53 ± 3.56	<0.001	28.49 ± 4.70	27.81 ± 3.93	28.67 ± 4.07	<0.001
Maternal age ≥35 (%)	8311 (13.11)	5734 (15.35)	525 (8.95)	97 (8.19)	<0.001	1880 (10.68)	59 (5.20)	16 (5.86)	<0.001
Multipara (%)	30,463 (48.07)	20,537 (55.00)	624 (10.64)	28 (2.36)	<0.001	9230 (52.48)	38 (3.36)	6 (2.20)	<0.001
Gestational weeks	38.97 ± 1.04	38.95 ± 1.00	39.08 ± 0.98	39.15 ± 1.00	<0.001	38.94 ± 1.11	39.18 ± 1.15	39.11 ± 1.15	<0.001
ART (%)	1372 (2.16)	755 (2.02)	207 (3.53)	87 (7.35)	<0.001	260 (1.48)	45 (3.97)	18 (6.59)	<0.001
Comorbidities (%)									
GDM or PGDM	9715 (15.32)	6216 (16.65)	1038 (17.70)	223 (18.83)	0.024	2064 (11.73)	126 (11.11)	48 (17.58)	0.009
Hypertensive disorder	3355 (5.29)	1897 (5.08)	304 (5.18)	46 (3.89)	0.164	1004 (5.70)	82 (7.23)	22 (8.06)	0.030
Uterine myoma	1700 (2.68)	1074 (2.88)	207 (3.53)	59 (4.98)	<0.001	320 (1.82)	34 (3.00)	6 (2.20)	0.017
Thyroid disorder	5802 (9.15)	2383 (6.38)	422 (7.20)	90 (7.60)	0.020	2562 (14.55)	255 (22.49)	90 (32.97)	<0.001
Length of labor stage (min)									
First stage	439.47 ± 258.33	403.21 ± 235.17	588.48 ± 260.76	660.13 ± 294.74	<0.001	424.68 ± 254.32	767.88 ± 297.38	830.57 ± 276.47	<0.001
Second stage	31.68 ± 29.55	22.85 ± 14.26	80.78 ± 16.07	149.17 ± 27.18	<0.001	21.15 ± 12.83	80.56 ± 16.49	151.44 ± 28.16	<0.001
Third stage	6.58 ± 6.62	6.39 ± 6.33	6.44 ± 6.73	8.28 ± 7.66	<0.001	6.88 ± 7.18	6.73 ± 4.49	6.61 ± 3.73	0.642
Total stage	477.62 ± 273.45	431.02 ± 241.63	676.8 ± 266.75	815.43 ± 301.19	<0.001	452.38 ± 260.4	853.85 ± 300.63	988.89 ± 283.77	<0.001
Related features (%)									
PROM	12,863 (20.29)	7600 (20.35)	1546 (26.36)	352 (29.73)	<0.001	3025 (17.18)	266 (23.46)	74 (27.11)	<0.001
Epidural analgesia	19,097 (30.12)	13,619 (36.47)	2948 (50.27)	636 (53.72)	<0.001	1398 (7.94)	363 (32.01)	133 (48.72)	<0.001
Artificial rupture of membranes	14,563 (22.97)	7998 (21.42)	1493 (25.46)	374 (31.59)	<0.001	4128 (23.45)	453 (39.95)	117 (42.86)	<0.001
Non-OA fetal position	5264 (8.30)	709 (1.90)	236 (4.02)	74 (6.25)	<0.001	4075 (23.15)	158 (13.93)	12 (4.40)	<0.001
Polyhydramnios	755 (1.19)	575 (1.54)	107 (1.82)	22 (1.86)	0.201	50 (0.28)	0 (0)	1 (0.37)	0.107
Oligohydramnios	1278 (2.02)	960 (2.57)	126 (2.15)	21 (1.77)	0.043	162 (0.92)	7 (0.62)	2 (0.73)	0.621
Macrosomia	1327 (2.09)	700 (1.87)	111 (1.89)	24 (2.03)	0.928	450 (2.56)	34 (3.00)	8 (2.93)	0.617
Low birthweight infants	1321 (2.08)	826 (2.21)	87 (1.48)	10 (0.84)	<0.001	386 (2.19)	10 (0.88)	2 (0.73)	0.001
Maternal and neonatal outcomes (%)									
Intrauterine infection	419 (0.66)	261 (0.70)	85 (1.45)	59 (4.98)	<0.001	13 (0.07)	1 (0.09)	0 (0)	0.659
PPH	2991 (4.72)	1719 (4.60)	399 (6.80)	105 (8.87)	<0.001	644 (3.66)	92 (8.11)	32 (11.72)	<0.001
Male fetus	32,948 (51.97)	19,182 (51.37)	3065 (52.27)	628 (53.04)	0.252	9311 (52.89)	623 (54.94)	140 (51.28)	0.348
Birthweight (kg)	3.21 ± 0.37	3.20 ± 0.37	3.23 ± 0.35	3.27 ± 0.34	<0.001	3.22 ± 0.38	3.28 ± 0.35	3.30 ± 0.34	<0.001
Neonatal asphyxia[a]	277 (0.44)	130 (0.35)	36 (0.61)	24 (2.03)	<0.001	67 (0.38)	16 (1.41)	4 (1.47)	<0.001
NICU admission	4617 (7.28)	2059 (5.51)	456 (7.78)	201 (16.98)	<0.001	1715 (9.74)	150 (13.23)	36 (13.19)	<0.001

ART, Assistant reproduction technology; GDM, Gestational diabetes mellitus; PGDM, pregestational diabetes mellitus; PROM, premature rupture of membranes; Non-OA, non-Occipital Anterior fetal position; PPH, postpartum hemorrhage; NICU, neonatal intensive care unit; 1 h, 1 hour; 1–2 h, 1–2 hours; 2 h, 2 hours. Data were presented as mean ± SD or N (%). [a]Neonatal asphyxia: neonatal 1-min Apgar score was equal or less than 7.

Table 1: **Baseline characteristics of the study cohort.**

GB model (0.824 [95% CI 0.804–0.843] for testing set; 0.859 [95% CI 0.843–0.875] for validation set) was significantly higher than other models (Fig. 3C and D). In the calibration plot, the GB model showed the best calibration among the ML models (Supplemental Figure S3). According to these results, GB algorithm was chosen to setup an online calculator (http://laborprediction.online) to predict the duration of second stage.

SHAP analysis was used to quantify the effect of each clinical feature on the prediction of the GB models. All eight features had a positive impact on predicting second-stage duration, as indicated by the clustering of red points at positive SHAP values (Fig. 4). To demonstrate how the model analyzes predictions, we selected two cases with similar clinical characteristics from the testing set to show how the model can be applied and interpreted using SHAP values (Fig. 5). Both cases involved 29-year-old nulliparous women at 39 weeks of gestation in the occiput anterior (OA) fetal position with epidural analgesia. The only difference between them was the duration of first stage. Woman No.28330, with a second-stage duration of 11 min, was predicted to be in the <1 h group by the 1 h model. Woman No.10121, with a second-stage duration of 120 min, was predicted to be in the ≥1 h group by the 1 h model.

For FAH cohort, the baseline data and results of feature selection are shown in the Supplemental Tabel S5,S9. Features were added step-by-step in the model development, and the model performances were

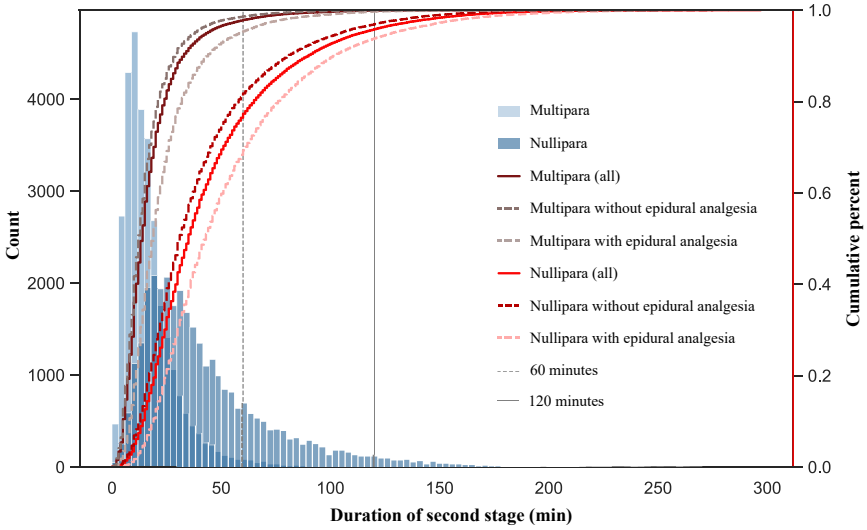

***Fig. 2*: Distribution of duration of second stage (min) across the study cohort (N = 63,401).** Red lines and right axis show the cumulative distribution of multipara, nullipara, respectively. Dashed and solid gray lines indicate the duration of second stage at 60 min and 120 min.

Features	1 h model					2 h model				
	≥1 h group	<1 h group	OR	95% CI	*P*	≥2 h group	<2 h group	OR	95% CI	*P*
N	7048 (15.88)	37,343 (84.12)				1184 (2.67)	43,207 (97.33)			
Maternal age (year)[a]	29.37 ± 3.77	30.05 ± 4.38	1.07	1.06–1.08	<0.001	29.53 ± 3.56	29.95 ± 4.32	1.09	1.07–1.10	<0.001
Nullipara	6396 (90.75)	16,806 (45.00)	11.99	11.03–13.03	<0.001	1178 (96.80)	23,167 (50.46)	39.66	27.25–57.72	<0.001
Gestational weeks	39.09 ± 0.99	38.95 ± 1.00	1.15	1.12–1.18	<0.001	39.15 ± 1.00	38.97 ± 1.00	1.20	1.13–1.28	<0.001
Duration of first stage (hour)	10.00 ± 4.47	6.72 ± 3.92	1.18	1.18–1.19	<0.001	11.00 ± 4.91	7.14 ± 4.12	1.18	1.17–1.19	<0.001
PROM	1898 (26.93)	7600 (20.35)	1.44	1.36–1.53	<0.001	352 (29.73)	9146 (21.17)	1.58	1.39–1.79	<0.001
Epidural analgesia	3584 (50.85)	13,619 (36.47)	1.80	1.71–1.9	<0.001	636 (53.72)	16,567 (38.34)	1.87	1.66–2.10	<0.001
Non-OA fetal position	310 (4.40)	709 (1.90)	2.38	2.08–2.72	<0.001	74 (6.25)	945 (2.19)	2.98	2.34–3.81	<0.001
Artificial rupture of membrane	1867 (26.49)	7998 (21.42)	1.32	1.25–1.40	<0.001	374 (31.59)	9491 (21.97)	1.64	1.45–1.86	<0.001

PROM, premature rupture of membranes; non-OA, non-Occipital Anterior fetal position; OR, odds ratio; CI, confidence interval; RFE, recursive feature elimination; 1 h, 1 hour; 2 h, 2 hours. Data were presented as mean ± SD or N (%). [a]Adjusted by parity.

***Table 2*: The features selected by RFE in model derivation (N = 44,391).**

	Internal validation							External validation						
	AUC (95% CI)	Acc	Sen	Spe	PPV	NPV	F1 score	AUC (95% CI)	Acc	Sen	Spe	PPV	NPV	F1 score
1 h models														
GB	0.808 (0.797–0.819)	0.676	0.847	0.643	0.311	0.957	0.455	0.862 (0.854–0.870)	0.645	0.937	0.622	0.165	0.992	0.281
RF	0.800 (0.790–0.811)[a]	0.667	0.847	0.633	0.306	0.956	0.449	0.858 (0.850–0.866)[a]	0.639	0.935	0.615	0.163	0.992	0.277
KNN	0.786 (0.775–0.797)[a]	0.670	0.835	0.638	0.306	0.953	0.448	0.834 (0.826–0.843)[a]	0.648	0.907	0.628	0.163	0.988	0.276
GNB	0.790 (0.778–0.801)[a]	0.686	0.800	0.665	0.313	0.946	0.449	0.835 (0.826–0.844)[a]	0.630	0.920	0.607	0.158	0.990	0.269
2 h models														
GB	0.824 (0.804–0.843)	0.626	0.868	0.619	0.058	0.994	0.109	0.859 (0.843–0.875)	0.634	0.927	0.630	0.035	0.998	0.068
RF	0.814 (0.794–0.834)	0.633	0.863	0.626	0.059	0.994	0.110	0.842 (0.825–0.859)[a]	0.638	0.908	0.635	0.035	0.998	0.067
KNN	0.795 (0.774–0.816)[a]	0.610	0.859	0.603	0.055	0.994	0.104	0.836 (0.819–0.854)[a]	0.627	0.894	0.623	0.033	0.998	0.064
GNB	0.809 (0.788–0.831)[a]	0.586	0.889	0.578	0.054	0.995	0.102	0.824 (0.806–0.841)[a]	0.562	0.945	0.556	0.030	0.999	0.058

GB, gradient boosting machine implemented by decision tree; RF, random forest; KNN, K- nearest neighbor; GNB, Gaussian Naive Bayes; AUC, area under the receiver operator characteristic curve; CI, confidence interval; Acc, accuracy; Sen, sensitivity; Spe, specificity; PPV, positive predictive value; NPV, negative predictive value; 1 h, 1 hour; 2 h, 2 hours. [a]P value of DeLong's test compared with GB model was <0.05.

***Table 3*: Comparison of performance among ML models.**

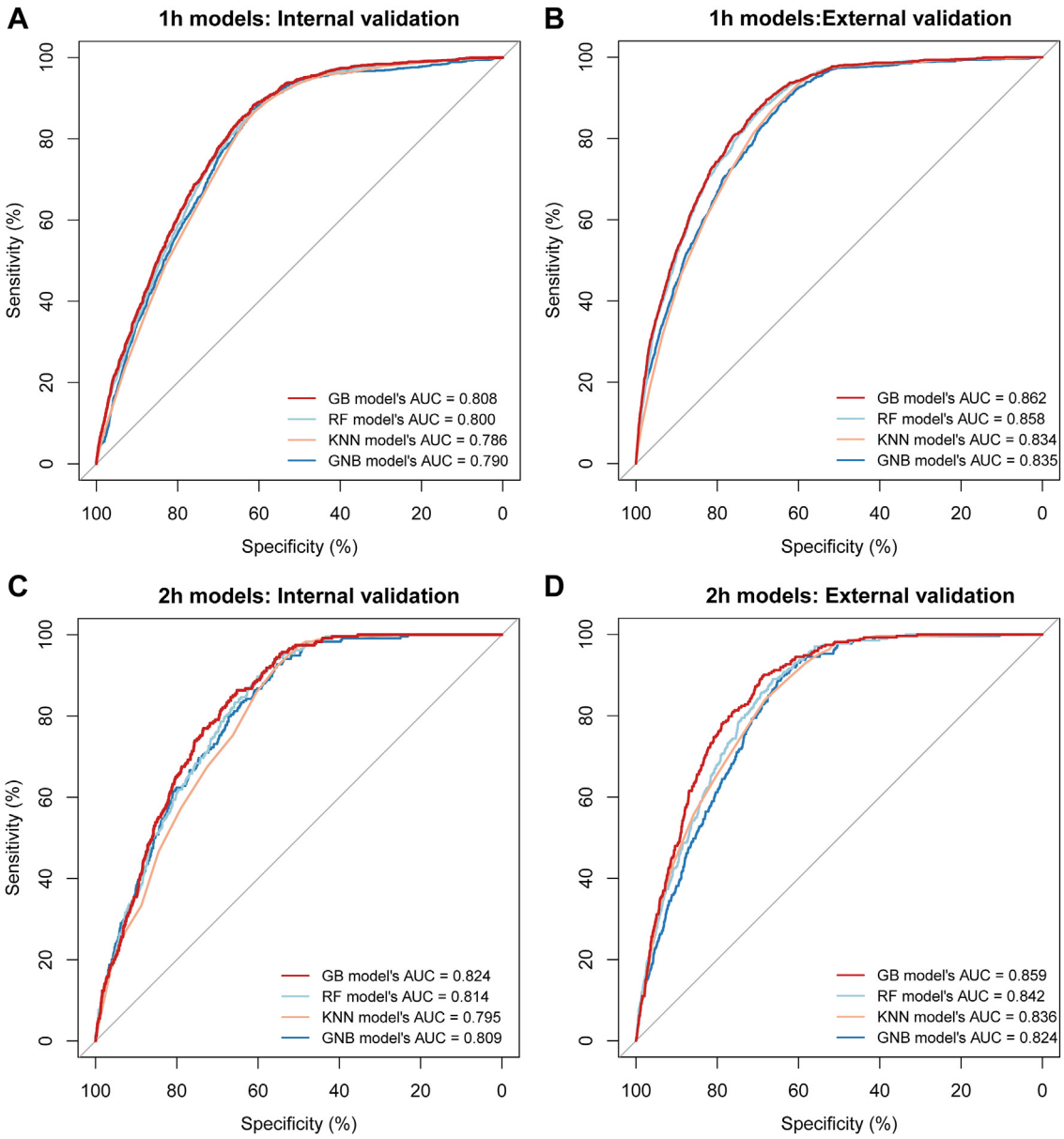

***Fig. 3*: ROC curves of ML models in internal and external validation**. (A, B) ROC curves of 1 h model; (C, D) ROC curves of 2 h model. Abbreviation: ROC, receiver operating characteristic; ML, machine learning; AUC, area under the curve; GB, gradient boosting machine implemented by decision tree; RF, random forest; KNN, K-nearest neighbor; GNB, Gaussian Naive Bayes; 1 h, 1 hour; 2 h, 2 hours.

gradually boosted as presented in Table 4. Compared to the crude model (Model 1) with the original eight features yielding an AUC of 0.816 (95% CI 0.794–0.839), the AUCs for Model 2 and Model 3 were 0.821 (95% CI 0.799–0.843) and 0.830 (95% CI 0.808–0.851), respectively. Model 4, which included all additional features, demonstrated the highest AUC (0.864, 95% CI 0.846–0.882), accuracy (0.751), sensitivity (0.858), specificity (0.727), and F1 score (0.562). The ROCs of the models were presented in Supplemental Figure S4. This model showed that adding clinical features, especially caput succedaneum, significantly improved the prediction of second-stage duration.

Discussion

The objective of the present study was to predict second-stage labor duration using artificial intelligence-based models. Based on the information of 63,401 singleton vaginal deliveries, this study has developed explainable ML models to predict second-stage labor duration.

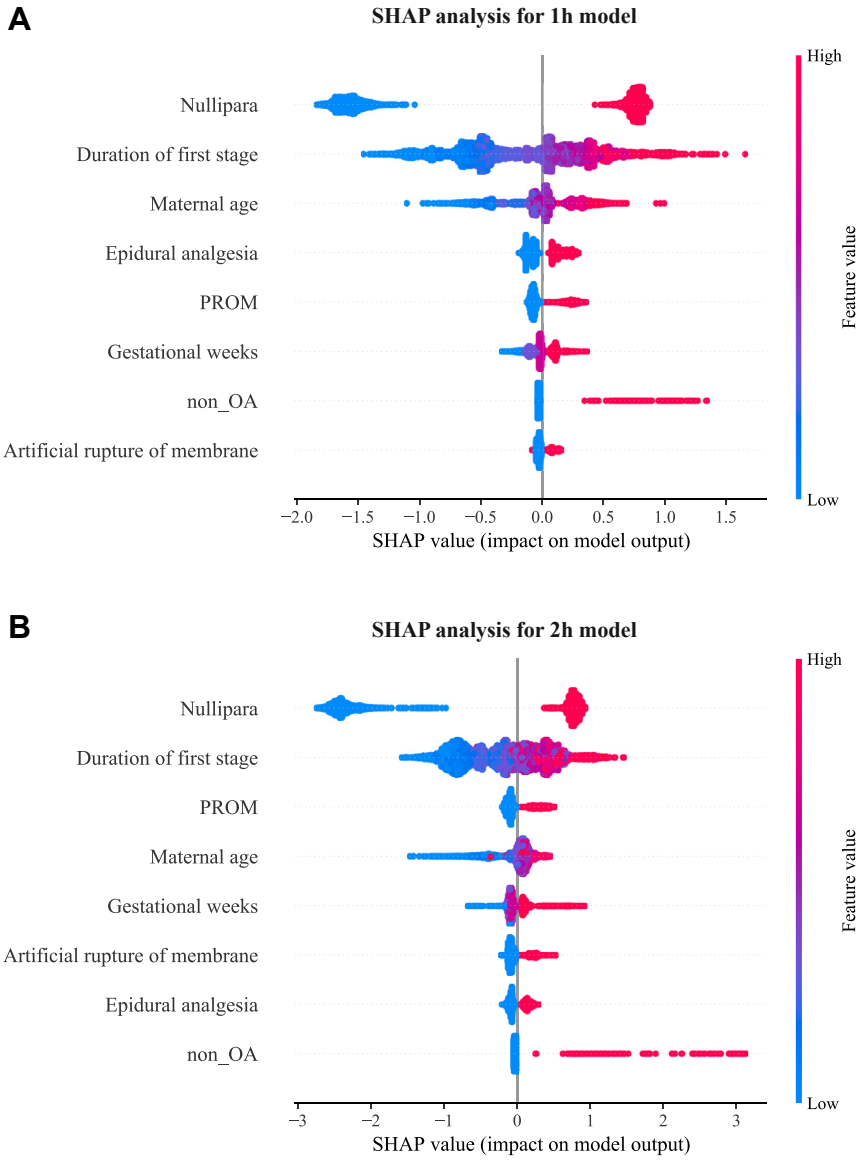

Fig. 4: **SHAP analysis for models of GB algorithm**. Global explanation of GB model with the mean of SHAP values, ranked by the importance of each feature. Each point color encodes the SHAP value of each clinical feature for each individual, red and blue colors for high and low values of the feature, respectively. On the X-axis, a positive or negative SHAP value means that the feature contributed to a positive or negative estimation of the prediction result of duration of second stage of labor. A) SHAP analysis for 1 h model; B) SHAP analysis for 2 h model; Abbreviation: SHAP, Shapley Additive exPlanations; GB, gradient boosting machine implemented by decision tree; PROM, premature rupture of membranes; non-OA, non-Occipital Anterior fetal position; 1 h, 1 hour; 2 h, 2 hours.

Prolonged second stage of delivery increased perinatal risks, including operative vaginal deliveries, severe perineal lacerations, PPH, and neonatal-fetal hypoxia.[1,4,5,7,37] Both global[10] and regional clinical practice guidelines[9,13,14,16,38] all suggest the upper time limit of second stage of labor. The American College of Obstetricians and Gynecologists (ACOG) guideline[12] and the National Institute for Health and Care Excellence (NICE) guideline[15] both suggest an individual active second stage. However, there is no precise method to predict the second stage of labor, currently. The present ML models could predict individual second-stage duration with an AUC up to 0.824. With the help of these models, obstetricians and midwives can identify women with a longer second stage and provide appropriate clinical intervention.

Notably, the models only consist of non-invasive and easily-accessed clinical information. The impacts of

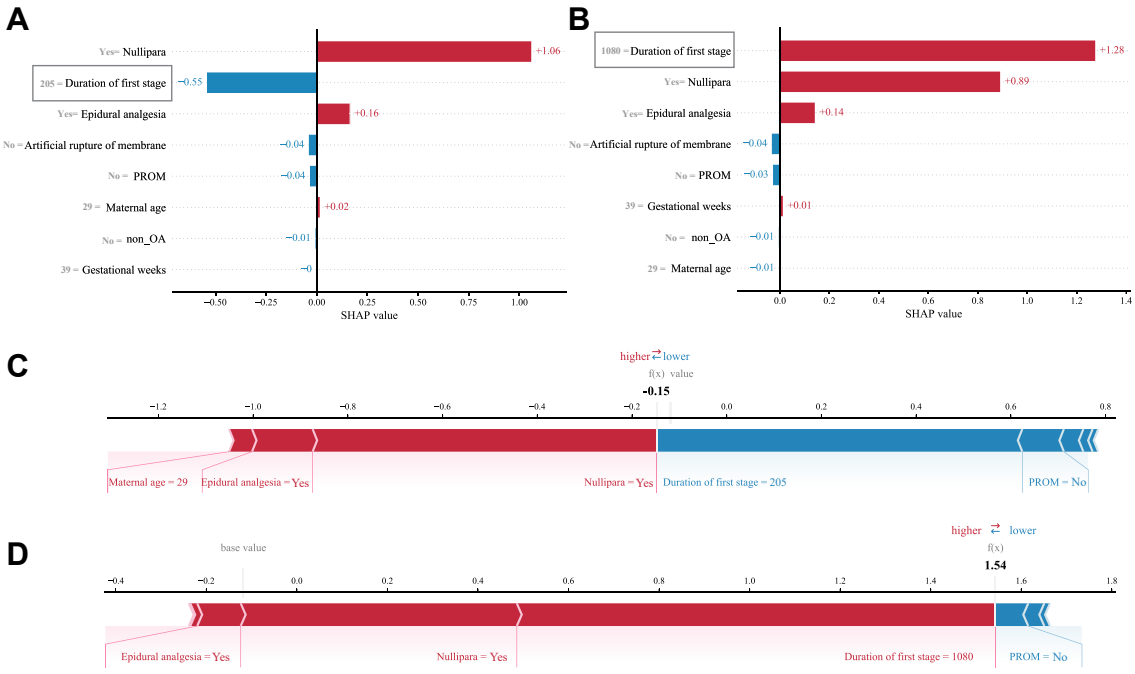

***Fig. 5*: Two similar examples for GB model prediction with SHAP analysis**. Importance plots and force plots visualized individual model prediction as result of feature contributions, showing the process how the model making individual decision. Red and blue colors presented as positive and negative SHAP value, respectively. Details information of each woman showed in gray to the left of the feature in A and B, where the different features between them were framed. (A, C) A 29-year-old nulliparous woman with second-stage duration of 11 min, was exactly predicted as <1 h Group by 1 h model. (B, D) A 29-year-old nulliparous woman with second-stage duration of 120 min, was exactly predicted as ≥1 h Group by 1 h model. Abbreviation: SHAP, Shapley Additive exPlanations; GB, gradient boosting machine implemented by decision tree; PROM, premature rupture of membranes; non-OA, non-Occipital Anterior fetal position; 1 h, 1 hour.

parity, epidural analgesia, and premature or artificial rupture of membrane on the duration of second stage were well recognized.[9,11,33] We also identified that longer first stage and elder maternal age were both associated with prolonged second stage, which is consistent with previous reports.[39–42] A possible explanation for this might be the myometrium is less effective to oxytocin as aging, and a recent single-cell study also found myometrial aging characteristics related to pregnancy and labor.[43] These results indicated that both prolonged first and second stage may be associated with maternal age.

Another strength of the present study is the use of ML algorithms to quantify and integrate clinical information. ML algorithms demonstrate significant advantages in processing multi-dimensional large-scale datasets.[44,45] Guedalia et al. used ML model to predict successful vaginal delivery,[20] and Gimovsky AC et al. also applied ML model to predict spontaneous vaginal delivery.[7] The ML model reported in the present study can precisely classified second-stage duration among diverse clinical conditions, including spontaneous and induced labor, and many types of complications. Moreover, SHAP analysis provided interpretable and fair explanations to the models, making them explainable in clinical applications. To our knowledge, this is the first study to predict the duration of second stage of labor using artificial intelligence methods.

This study also has some limitations. Firstly, some potential parameters were not available because of the retrospective design, including estimated fetal weight

	Model 1	Model 2	Model 3	Model 4
AUC	0.816	0.821	0.830	0.864
95% CI	0.794–0.839	0.799–0.843	0.808–0.851	0.846–0.882
P value[a]	reference	0.004	<0.001	<0.001
Accuracy	0.696	0.700	0.713	0.751
Sensitivity	0.786	0.778	0.789	0.858
Specificity	0.675	0.682	0.695	0.727
PPV	0.357	0.359	0.372	0.418
NPV	0.933	0.931	0.935	0.957
F1 score	0.491	0.491	0.506	0.562

AUC, the area under the receiver operator characteristic curve; CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value; FAH, The First Affiliated Hospital of Sun Yat-sen University. Model 1: the crude model with the same eight features of Table 2. Model 2: Model 1 + maternal weight. Model 3: Model 2 + oxytocin augmentation. Model 4: Model 3 + caput succedaneum. [a]*P* value of DeLong's test.

***Table 4*: Comparison of the performance of different models in FAH cohort.**

and the use of oxytocin. The precision of estimated fetal weight (EFW) various among cases due to the difference of ultrasound examiners and the time gap between the date of examination and delivery, and the use of oxytocin also various among different physicians. A further limitation is the absence of models to assess second-stage labor durations exceeding 3 h and 4 h, primarily due to the insufficient sample size in these categories. Nonetheless, prediction and management of second-stage labor exceeding 1 h and 2 h remain important to those with maternal or fetal complications, such as fetal distress, maternal heart disease, etc. Thirdly, the current model is applicable only to spontaneous deliveries and does not extend to cases of operative delivery. This limitation arises from the diverse underlying reasons for operative delivery during labor, making it difficult to build a predictive model. In addition, the accuracy of model was slightly decreased in the external validation. Because the validation set is geographically independent to the derivation set, the difference of their clinical features and outcome rates may contribute to model drift between training and validation sets.

In the future, well-designed prospective cohorts from multi-ethnic populations would help to achieve better implications and utilization of artificial-intelligence based models. In addition, more clinical and laboratory information could be included to improve the models. Moreover, deep learning algorithms could be employed to create real-time multimodal models.

The present study was designed to create individual predictive models for second-stage labor duration with ML algorithms. This study has shown that the explainable ML models with quantified non-invasive clinical features, could serve as valuable tools to predict second-stage labor duration. These models have implications within the clinical setting for individualized labor management to improve maternal and neonatal outcomes.

Contributors

Bin Liu designed and organized the study. Xiaoqing Huang, Xiaodan Di, and Suiwen Lin were responsible for data collection and statistics analysis, involved in technique support in machine learning method with Bin Liu. Xiaoqing Huang drafted the manuscript. Xiaoqing Huang, Suiwen Lin and Bin Liu assessed and verified the data. Minrong Yao, Suijin Zheng, Shuyi Liu, Wayan Lau, Zhixin Ye, and Zilian Wang contributed to acquisition, analysis, or interpretation of data. Bin Liu critically reviewed the manuscript. All authors were involved in reviewing the manuscript and approved the final manuscript for submission.

Data sharing statement

Data are available on request from the corresponding author upon reasonable request. The python code and statistical analysis methods used in the study can be requested directly from the corresponding author after approval.

Declaration of interests

All authors report no conflict of interest related to this work.

Acknowledgements

This research was supported by National Natural Science Foundation of China (No.82371689, N0.81771602) and National Key Research and Development Program of China (No.2021YFC2700703). We acknowledge Professor Jinxin Zhang, and Dr. Shuo Yang from the Department of Medical Statistics, School of Public Health, Sun Yat-sen University for their contribution on statistical analysis.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.eclinm.2025.103072.

References

1 He X, Zeng X, Troendle J, et al. New insights on labor progression: a systematic review. *Am J Obstet Gynecol*. 2023;228(5s):S1063–S1094.
2 Friedman EA. Primigravid labor; a graphicostatistical analysis. *Obstet Gynecol*. 1955;6(6):567–589.
3 Zhang J, Landy HJ, Ware BD, et al. Contemporary patterns of spontaneous labor with normal neonatal outcomes. *Obstet Gynecol*. 2010;116(6):1281–1287.
4 McKinney JR, Allshouse AA, Heyborne KD. Duration of labor and maternal and neonatal morbidity. *Am J Obstet Gynecol MFM*. 2019;1(3):100032.
5 Grantz KL, Sundaram R, Ma L, et al. Reassessing the duration of the second stage of labor in relation to maternal and neonatal morbidity. *Obstet Gynecol*. 2018;131(2):345–353.
6 Kleinstern G, Zigron R, Porat S, et al. Duration of the second stage of labour and risk of subsequent spontaneous preterm birth. *BJOG An Int J Obstet Gynaecol*. 2022;129(10):1743–1749.
7 Gimovsky AC, Levine JT, Pham A, Dunn J, Zhou D, Peaceman AM. Pushing the bounds of second stage in term nulliparas with a predictive model. *Am J Obstet Gynecol MFM*. 2019;1(3):100028.
8 Young C, Bhattacharya S, Woolner A, et al. Maternal and perinatal outcomes of prolonged second stage of labour: a historical cohort study of over 51,000 women. *BMC Pregnancy Childbirth*. 2023;23(1):467.
9 [Guideline of normal birth]. *Zhonghua Fu Chan Ke Za Zhi*. 2020;55(6):361–370.
10 WHO Guidelines Approved by the Guidelines Review Committee. *WHO recommendations: intrapartum care for a positive childbirth experience*. Geneva: World Health Organization Copyright © World Health Organization 2018; 2018.
11 Caughey AB, Cahill AG, Guise J-M, Rouse DJ. Safe prevention of the primary cesarean delivery. *Am J Obstet Gynecol*. 2014;210(3):179–193.
12 First and Second Stage Labor Management. ACOG clinical practice guideline No. 8. *Obstet Gynecol*. 2024;143(1):144–162.
13 Queensland Health. Maternity and neonatal clinical guideline: normal birth. (MN22.25-V5-R27)2022. https://www.health.qld.gov.au/__data/assets/pdf_file/0014/142007/g-normalbirth.pdf. Accessed July 30, 2024.
14 Royal Australian and New Zealand College of Obstetricians and Gynaecologists. Care in labour in the absence of pregnancy complications (C-Obs 31)2023. https://ranzcog.edu.au/wp-content/uploads/2022/05/Care-in-Labour-in-the-Absence-of-Pregnancy-Complications-C-Obs-31.pdf. Accessed July 30, 2024.
15 National Institute for Health and Care Excellence. *Clinical guidelines. Intrapartum care*. London: National Institute for Health and Care Excellence (NICE) Copyright © NICE 2023; 2023.
16 Lee L, Dy J, Azzam H. Management of spontaneous labour at term in healthy women. *J Obstet Gynaecol Can*. 2016;38(9):843–865.
17 Gimovsky AC, Berghella V. Randomized controlled trial of prolonged second stage: extending the time limit vs usual guidelines. *Am J Obstet Gynecol*. 2016;214(3):361.e1–361.e6.
18 Leveno KJ, Nelson DB, McIntire DD. Second-stage labor: how long is too long? *Am J Obstet Gynecol*. 2016;214(4):484–489.
19 Nelson DB, McIntire DD, Leveno KJ. Second-stage labor: consensus versus science. *Am J Obstet Gynecol*. 2020;222(2):144–149.
20 Guedalia J, Lipschuetz M, Novoselsky-Persky M, et al. Real-time data analysis using a machine learning model significantly improves prediction of successful vaginal deliveries. *Am J Obstet Gynecol*. 2020;223(3):437.e1–437.e15.
21 Bukowski R, Schulz K, Gaither K, et al. Computational medicine, present and the future: obstetrics and gynecology perspective. *Am J Obstet Gynecol*. 2021;224(1):16–34.

22 Lipschuetz M, Guedalia J, Rottenstreich A, et al. Prediction of vaginal birth after cesarean deliveries using machine learning. *Am J Obstet Gynecol*. 2020;222(6):613.e1–613.e12.
23 Guedalia J, Lipschuetz M, Cohen SM, et al. Transporting an artificial intelligence model to predict emergency cesarean delivery: overcoming challenges posed by interfacility variation. *J Med Internet Res*. 2021;23(12):e28120.
24 Westcott JM, Hughes F, Liu W, Grivainis M, Hoskins I, Fenyo D. Prediction of maternal hemorrhage using machine learning: retrospective cohort study. *J Med Internet Res*. 2022;24(7):e34108.
25 Dhombres F, Bonnard J, Bailly K, Maurice P, Papageorghiou AT, Jouannic JM. Contributions of artificial intelligence reported in obstetrics and gynecology journals: systematic review. *J Med Internet Res*. 2022;24(4):e35465.
26 Guedalia J, Sompolinsky Y, Novoselsky Persky M, et al. Prediction of severe adverse neonatal outcomes at the second stage of labour using machine learning: a retrospective cohort study. *BJOG An Int J Obstet Gynaecol*. 2021;128(11):1824–1832.
27 Schmidt LJ, Rieger O, Neznansky M, et al. A machine-learning-based algorithm improves prediction of preeclampsia-associated adverse outcomes. *Am J Obstet Gynecol*. 2022;227(1):77.e1–77.e30.
28 Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322(18):1806–1816.
29 Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med*. 2015;162(1):W1–W73.
30 Qi H, Yang H, Duan T. [The concern and adoption of the new standard of normal and abnormal labor]. *Zhonghua Fu Chan Ke Za Zhi*. 2014;49(7):487–489.
31 Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics*. 2001;17(6):520–525.
32 May JA, Feng Z, Adamowicz SJ. A real data-driven simulation strategy to select an imputation method for mixed-type trait data. *PLoS Comput Biol*. 2023;19(3):e1010154.
33 Piper JM, Bolling DR, Newton ER. The second stage of labor: factors influencing duration. *Am J Obstet Gynecol*. 1991;165(4 Pt 1):976–979.
34 Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn*. 2002;46(1):389–422.
35 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(null):2825–2830.
36 Lundberg SM, Lee S-I. *A unified approach to interpreting model predictions. Proceedings of the 31st international conference on neural information processing systems*. Long Beach, California, USA: Curran Associates Inc.; 2017:4768–4777.
37 Finnegan CL, Burke N, Breathnach F, et al. Defining the upper limit of the second stage of labor in nulliparous patients. *Am J Obstet Gynecol MFM*. 2019;1(3):100029.
38 Ushida T, Matsuo S, Nakamura N, et al. Reassessing the duration of each stage of labor and their relation to postpartum hemorrhage in the current Japanese population. *J Obstet Gynaecol Res*. 2022;48(7):1760–1767.
39 Greenberg MB, Cheng YW, Sullivan M, Norton ME, Hopkins LM, Caughey AB. Does length of labor vary by maternal age? *Am J Obstet Gynecol*. 2007;197(4):428.e1–428.e7.
40 Zaki MN, Hibbard JU, Kominiarek MA. Contemporary labor patterns and maternal age. *Obstet Gynecol*. 2013;122(5):1018–1024.
41 Tilden EL, Snowden JM, Bovbjerg ML, et al. The duration of spontaneous active and pushing phases of labour among 75,243 US women when intervention is minimal: a prospective, observational cohort study. *eClinicalMedicine*. 2022;48:101447.
42 Nelson DB, McIntire DD, Leveno KJ. Relationship of the length of the first stage of labor to the length of the second stage. *Obstet Gynecol*. 2013;122(1):27–32.
43 Punzon-Jimenez P, Machado-Lopez A, Perez-Moraga R, et al. Effect of aging on the human myometrium at single-cell resolution. *Nat Commun*. 2024;15(1):945.
44 Deo RC. Machine learning in medicine. *Circulation*. 2015;132(20):1920–1930.
45 Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4):233–234.