npj Digital Medicine

https://doi.org/10.1038/s41746-026-02384-0

Article in Press

External validation of a machine learning model for delivery mode prediction after induction

Received: 11 October 2025

Accepted: 15 January 2026

Cite this article as: Ferreira, I., Simões, J., Correia, J. *et al.* External validation of a machine learning model for delivery mode prediction after induction. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-02384-0

Iolanda Ferreira, Joana Simões, João Correia & Ana Luísa Areia

Article title: **External validation of a machine learning model for delivery mode prediction after induction**

Authors:

Author 1: Iolanda Ferreira MD [1,2],

Author 2: Joana Simões MSc [3],

Author 3: João Correia PhD [3],

Author 4: Ana Luísa Areia PhD [1,2,4]

Affiliations*:*

Obstetrics Department, Unidade Local de Saúde Coimbra, Coimbra, Portugal [1];

Faculty of Medicine of University of Coimbra, Coimbra, Portugal [2]

University of Coimbra, Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, Coimbra, Portugal [3]

ICBR Inflammation and Biomarkers Group, Coimbra, Portugal [4]

Corresponding author:

Name: Iolanda Ferreira

Email: 10862@chuc.min-saude.pt

Abstract:

No machine learning (ML) models for predicting delivery mode after labor induction (IOL) have been externally validated. We aimed to develop and validate one using medical records. Portuguese tertiary center data (n=2434) were used for development and internal validation, and Consortium on Safe Labor data (n=10591) for external validation. Outcomes are vaginal delivery (VD) or cesarean section (CS). Internal validation employed different ML approaches, aiming for model simplification. Logistic regression performed best on internal validation: AUROC:0.793; F1-score:0.748; PPV:0.752, with good calibration and decision curve analysis (DCA), being selected for simplification. Simplified top-13 features model was selected for external validation: AUROC:0.808; F1-score:0.781; PPV:0.822, tending for VD (99.6%) while avoiding false-positives (0.5%). Calibration curves underestimated CS risk by 10-75%; DCA showed good net benefit. The model's good AUROC and DCA suggest clinical utility. Calibration curve underestimation of CS risk may result from outcome imbalance between datasets.

Keywords: Machine learning, prediction model, mode of delivery, labor induction, external validation

Key message: This large external validation study of a machine learning model for delivery mode prediction shows good performance (AUROC: 0.808) and good net benefit on decision curve analysis, indicating its potential clinical utility for future impact assessment studies.

INTRODUCTION

Labor induction is an obstetric medical intervention used performed for a range of clinical and elective reasons in both high- and low-risk pregnancies,, and its use has increased over the past decade.[1] Achieving a safe vaginal delivery (VD) is the ultimate goal of induction of labor (IOL). While low-risk nulliparous women do not have higher caesarean section (CS) rates when induced at 39 weeks, induction at or beyond term can improve maternal and neonatal outcomes in high-risk groups.[1-3] Hence, though CS rates in low-risk nulliparous patients undergoing IOL are around 19%, 30% rates have been reported.[4] Considering that time until delivery may be increased in IOL compared to spontaneous onset of labor, and that several other factors could precipitate an unnecessary CS decision in this context, it could result in more extended hospital stays, labor medicalization, and growing health personnel needs and costs. [5,6] Although induction of labor is offered to all eligible women, current practice largely follows standardized, guideline-based protocols that prioritize maternal or fetal risk categories rather than finer-grained clinical or ultrasonographic details. Thus, while obstetricians rely on these protocols and on their own clinical judgment to anticipate the likelihood of VD or CS during induction, the ability to predict outcomes for an individual patient remains limited.[7] Therefore, we have hypothesized that ML models could provide individualized risk estimates of cesarean delivery before IOL begins.

Recent literature suggests that Machine Learning (ML) can enhance the accuracy of prediction models by addressing nonlinear relationships between variables. [6,8] Studies have used ML for delivery mode prediction, but most combine spontaneous and induced labor.[9] Few use datasets including only induced labor[10,11], and have limitations on sample size and inclusion criteria, particularly regarding induction methods, Bishop scores, and unscarred uterus. Recently, *Ferreira et al.* developed an ML model to predict delivery mode after IOL, achieving better results than prior studies.[12] Moreover, no ML model has been externally validated, limiting their potential clinical use. Creating an ML solution to assist pregnant patients and clinicians with delivery mode counseling before IOL has become crucial research in Obstetrics because rising IOL rates affect millions of pregnant individuals worldwide.[7,13]

As such, our primary objective was to adapt an existing model[12] or predicting delivery mode and evaluate its transportability in a geographically different population using variables derived from electronic medical records before attempting IOL. Secondary aims involve evaluating feature

importance and model simplification for clinical use, focusing on significant variables without compromising predictive accuracy.

RESULTS

Patient demographics

The Portuguese dataset included 2672 singleton, term-induced births (14.8% of total deliveries), of which 2434 met the inclusion and exclusion criteria. Of the 228 438 women in the CSL database, 77971 underwent IOL. Following the application of the previously outlined inclusion and exclusion criteria, 10591 participants remained for analysis (See patient flow diagram in Supporting Information Figure S1). CS rates were significantly higher in the Portuguese dataset (28.7%; n=698 vs 16.4%; n=1736). For the most common CS indications, the Portuguese dataset also showed significantly higher rates for fetal heart rate abnormalities (8.0% vs 4.5%), labor dystocia (11.2% vs 0.4%), and failed induction (7.7% vs 0.1%).

Table 1 displays an overall comparison between datasets. In brief, they differed statistically in terms of age, parity, and history of previous CS, with a higher proportion of older (32.7±5.5 vs 28.1±5.4 years), nulliparous (58.2% vs 38.9%), and women with prior CS (11.3% vs 2.0%) in the Portuguese dataset. Bishop's score at admission was significantly higher in the CSL dataset, which might relate to the most common IOL method used in this context: 72.8% used oxytocin. In comparison, mainly dinoprostone was used in the Portuguese dataset (54.4%). The latter was most frequently induced due to post-date pregnancy (43.0%), while other IOL indications were more common in the CSL dataset (54.8%). For a comparison of demographic features by delivery mode, and information on IOL data and maternal and neonatal outcomes, see Supporting Information Table S2.

Model development and external validation

Of the various 43-feature models developed on internal validation, LR showed a higher average AUROC (0.793, 95% CI 0.792–0.794), F1-score (0.748, 95% CI 0.747-0.749), and PPV (0.752, 95% CI 0.751-0.753), thus being selected for additional evaluation (Table 2). It predicted on average 91.2% VD and 40.7% CS, corresponding to 8.9% FP and 59.3% false-negative rates (see CM in Supporting Information Figure S2). Subsequently, SHAP analysis (Figure 1) evaluated the

LR model´s feature importance. Parity was the most influential predictor, with higher values tending towards VD, followed by previous CS, favoring CS outcomes. The same trend was observed for increasing maternal age, final pregnancy weight, and gestational age at induction. Conversely, increasing Bishop score and maternal height favored VD. With features ranked by importance, the elbow method allowed model simplification (Supporting Information Figure S3). The LR model with the 13 top features on the SHAP plot (Table 2 and Supporting Information Figure S2) performed best and was selected for external validation. The test results on the CSL dataset indicate higher AUROC, F1-score, and PPV values (0.808, 0.781, and 0.822, respectively); see Figure 2 and Table 2. Its CM tended for VD classification (true negative (TN): 99.6%) at the expense of avoiding FP (0.5%), i.e., classifying true VD as CS. The true positive rate (correctly classifying CS) was low (5.7%), confirming this tendency (Supporting Information Figure S2).

As expected, the 13-variable LR model showed good calibration curves in the Portuguese dataset (intercept: 0.01 (95% CI 0.010–0.013); slope: 1.05 (95% CI 0.042–0.052). However, when tested in the CSL dataset, it consistently underestimated CS risk between 10-75%, where the graphs drop off sharply, indicating that CS risk is overestimated beyond this percentage (intercept: 0.26; slope: 0.63; Figure 3). This pattern is in line with the CM metrics described above. DCA also demonstrated that the model exceeds the net benefit of treating all versus none at every threshold on the Portuguese dataset, while this was observed between 10-75% CS risk threshold on the CSL dataset (Supporting Information Figure S4).

DISCUSSION

We report the adaptation and transportability of a previously reported ML model for delivery mode prediction after IOL.[12] Its simplified 13-variable approach showed good AUROC, calibration, and DCA values on the development (Portuguese) dataset. External validation metrics demonstrated superior AUROC values and stable DCA curves. However, model calibration was suboptimal, with a tendency for CS underestimation between 10-75% cut-off values, followed by an abrupt CS risk overestimation after this percentage. Feature importance evaluation was crucial in model simplification, potentially enhancing its ease of use in clinical practice and within the pregnant population.

After extensive research in several databases (using different combinations of the terms 'labor induction', 'machine learning', and 'external validation'), no studies report on external validation of delivery mode prediction models after IOL using ML. Accordingly, a recent review of ML applications in peripartum care reported external validation in only 5% of studies, with only 1% of the developed models leading to clinical tools.[14] External validation studies using classical statistical models for predicting delivery mode after IOL mostly report AUROC measures.[4,13,15,16] Among these studies, those using the CSL dataset report AUROC values similar to ours: 0.73 and 0.81 in the studies by *Kawakita et al*.[15] and *Jochum et al*.[13], respectively. *Kawakita et al´s* calibration curve is also similar to ours.[15] However, this is an incomplete comparison due to the different methodologies used by these authors and the lack of other calculations, such as DCA. In addition, only three ML prediction models for delivery mode after IOL have been developed so far. The most recent[12] presents an internally validated model yielding high-performance metrics: AUROC (0.794, 95% CI 0.783 – 0.805), and PPV values (0.752, 95% CI 0.751 – 0.753). Compared to the study by *D´Sousa et al*., it presents a better AUROC (0.79 vs 0.73) while using a larger dataset and proper validation methodology with cross-validation.[10] The study by *Hu et al.* also showed high-performance metrics, but their methodology lacked clarity on data preparation, analysis, and model validation to replicate and compare.[11]

Focusing on the 13-variable externally validated model, it showed good discrimination (AUROC 0.808). Although there is a significant outcome imbalance, F1-score and PPV still performed well (0.781 and 0.822, respectively), possibly due to participant numbers and the CSL dataset's diversity. However, likely due to the aforementioned outcome imbalance (CS rates of 28.7% on the Portuguese dataset vs 16.4% on the CSL dataset), this model showed some miscalibration.[17] This difference could be explained by different data collection periods, reflecting both a time lag of approximately 15 years and different clinical practices and population characteristics, which may limit the model's transportability. The CSL dataset includes younger, taller, mostly multiparous individuals with significantly better Bishop scores than the Portuguese dataset. It also presents lower previous CS rates and final pregnancy weight. Indeed, our feature importance analysis highlighted these features, all supported by literature, as clinically relevant for IOL.[10,13] Nonetheless, we opted to use the CSL dataset because it prompted several important publications, such as Zang's labor curves[18], as well as calculators that used statistical calculations on mode of delivery prediction after IOL.[7,15,19] Additionally, of the few available datasets that could have

offered a more contemporary approach to IOL, none included key information on IOL deliveries, such as induction indication or the method used. Nevertheless, we followed the correct methodology for model transportability evaluation, because when development and validation cohorts are very similar, only model reproducibility is assessed, not transportability.[17]

Finally, DCA demonstrated the model´s clinical net benefit in predicting CS risk between 10%-75%. Decision analysis frameworks evaluate the clinical trade-offs between two types of errors: FP (unnecessary CS) and false negative (missed CS). We chose our thresholds as "treat all" or "treat none," meaning we opted for the extremes of decision analysis to understand the model´s net benefit and its value as a decision tool, as supported by literature.[20] We interpreted our results by comparing these two extreme thresholds to define baseline strategies. Any model tested should have a net benefit higher than both thresholds across a reasonable range to be considered useful. This pattern was verified with internal and external validation results. However, before the model can be tested in a clinical impact assessment study, it would be relevant to update it better to suit the average outcome risk in the external population. Because an updated model is essentially a new model, it requires further internal and external validation before moving forward to clinical studies.[20] Our research group is currently collecting more data to create a dataset that will enable these subsequent steps.

When addressing studies concerning the development or validation of models for predicting birth mode, two topics are of primary concern. First, these models could influence CS decisions, particularly when assigning high CS probabilities, which could increase CS rates. Second, there is concern about their clinical utility because, ultimately, all women eligible for IOL deserve a trial of labor.

The model does not apply to the small sample of women who wish to undergo elective CS, but to all women desiring vaginal birth who wish counselling about their personalized mode of delivery probability after IOL. A discrete risk cut-off above which a cesarean is indicated was not defined because a woman's preference ultimately outweighs any model´s predictive percentages. As others have stated, these tools should be used alongside appropriate clinical context and judgement and are not prescriptive for CS when IOL seems less favorable, or rush CS decisions, as even women with lower success rates often labor successfully.[4] In this context, obstetrical guidelines should be followed for safe clinical labor management.

Therefore, the model could be helpful when there are clinical factors that might hasten CS decisions in the context of IOL, which might relate to medical or obstetric history or suspected macrosomia.[21] Recent studies have also shown low compliance with current ACOG guidelines on labor management, suggesting that clinicians consider other factors that affect dilation and descent.[6] Our algorithm could be reassuring in these scenarios, eventually preventing unnecessary CS decisions due to IOL failure, labor arrest, or cephalopelvic disproportion. However, we acknowledge CS rates may rise, as high-risk women may opt for CS even when they might ultimately achieve a VD.

While we cannot directly compare our models because they were developed on imbalanced datasets with different data quantities, both tend to avoid FP while maximizing the number of correctly predicted CS and VD. The 13-variable model´s CM highlights this, showing a tendency to predict VD, as obstetricians do in real clinical scenarios, to avoid performing CS in patients who would have VD. Therefore, it has a very low FP rate (0.45%) and a very high false-negative rate (94.3%). This tendency is also evident in the model's calibration curve, which overestimates VD at cut-off values between 10 and 75%, as well as in DCA, which provides a net benefit for clinical decision-making at the same thresholds. These metrics demonstrate the model's safety and suggest it could be tested in a clinical setting.

Given this pattern, the model functions primarily as a rule-out tool: it is extremely safe when predicting VD and can provide reassurance for proceeding with a standard trial of labor, but it is not reliable for identifying patients who will ultimately require CS. Its conservative nature aligns with real-world clinical decision-making but limits its usefulness for positive CS prediction. Recognizing these characteristics is essential for appropriate clinical interpretation and underscores the need for further calibration and prospective evaluation before recommending the model for real world clinical decision support.

The model's primary strength lies in using two distinct cohorts from different continents for separate model development and validation. These are also some of the largest and most comprehensive databases, including only deliveries after IOL from high- and low-risk pregnancies. These characteristics ensure the model´s transportability and fairness, making it applicable to a large IOL population.

Our study's retrospective nature is its most significant limitation. However, retrospective consultation of electronic medical records data could be integrated into software systems that incorporate predictive models like ours. In clinical practice, this system could provide real-time risk assessments for patients, aiding obstetricians in counseling about labor induction. It could identify women at low risk for cesarean, allowing for a standard trial of labor, while higher risk patients could receive closer monitoring. Integration into electronic health systems would enable continuous updates to improve its applicability and safety. We have also omitted data on race because this information was not included in the Portuguese dataset. Additionally, we acknowledge that predictive models may perform differently among various patient subgroups. Factors such as age, body mass index (BMI), induction indication and parity are significant when considering labor induction, and the model's predictions may vary across these different groups. Therefore, prospective validation in diverse cohorts to evaluate any potential algorithmic bias is essential to ensure fair application in clinical practice. Finally, though the Portuguese and CSL datasets are among the largest and most detailed IOL cohorts annotated for mode of delivery, they remain relatively small for more complex approaches. This limitation may affect the stability and generalizability of more advanced models and underscores the need for future validation in larger, contemporary datasets.

This study demonstrates the model has good AUROC and DCA metrics, performing best for intermediate risk levels, the grey area that clinicians and pregnant individuals must manage before IOL. There are no externally validated ML models on birth mode prediction after IOL. Of the studies employing classical statistics for external validation using the CSL dataset, only one demonstrates a complete calibration curve visually analogous to ours.[13] Despite calibration limitations, our study represents a substantial advancement in the sphere of ML predictive model external validation and may hold promise for future clinical applications. This scenario is possible because a model may be undercalibrated yet still possess adequate discrimination, indicating its capacity to accurately categorize outcomes without invalidating the significance of external validation. [17,22]

METHODS

Datasets

We developed and validated an ML prediction model using tabular maternal-fetal data from two independent datasets. The Portuguese dataset, previously described in[12], comprises a retrospective longitudinal cohort study conducted at a tertiary Portuguese Obstetrics center between January 2018 and December 2021, and was used to train, validate, and test the model (internal validation). External validation was performed with tabular data from the Consortium for Safe Labor (CSL) database. It was chosen to assess model performance across different geographic and temporal contexts because it includes induced labor and delivery information collected from electronic medical records from 19 United States hospitals from 2002 to 2008.[23] Inclusion criteria were the same for both datasets and included pregnant women ≥18 years of age with singleton pregnancies and a baseline Bishop score of ≤7, eligible for IOL. This rationale was used because Bishop scores over 8 indicate a VD probability similar to spontaneous labor.[24] Breech deliveries, planned CS, antepartum fetal demise, and fetal anomalies that prevent VD were also excluded from the study.

The model's input variables were supported by literature and collected on available electronic medical records from antenatal visits and IOL admissions.[24,25] They were assessed before IOL initiation and included participants' demographics, past and current medical and obstetric history, and clinical characteristics related to IOL (Supporting Information Table S1). Features collected during labor and after delivery were excluded from modeling to ensure predictions were based on IOL admission information. All participants underwent IOL according to ACOG and NICE recommendations.[24,25]

Dataset harmonization was performed by mapping and matching features from each dataset to ensure consistency, resulting in 43 matching features and an outcome class for the ground truth. Regarding data missingness, we opted to exclude eligible features if 60% or more of the values were missing.[26] Variables with a lower proportion were handled using simple imputation, using the median to impute numerical features and the mode (most frequent value) for categorical/binary variables. We chose univariate imputation for simplicity and to avoid the experimental caveats and potential overfitting of multivariate feature imputation. The overall missing data rates were 1.2% vs 1.6% and the maximum missing values per feature were 16.4% vs 21.1% (Portuguese and CSL datasets, respectively). As a result, no data was excluded.

Sample size

The minimum sample size was calculated accounting for 43 variables. It totaled a minimum of 1657 individuals with at least 464 events (CS) for the Portuguese dataset, and 3651 individuals with at least 585 events for the CSL dataset.[27] Since both datasets exceed the minimum sample size requirement (see Results section) and ML performance depends on data quantity and quality, we used all available data. The validity of this hypothesis is unaffected by feature selection for simplified modeling, as the number of features relates positively to sample size requirements.

The sample size calculation for the Portuguese dataset used a C-statistic from a previous prediction model study (AUROC 0.79)[12], which adjusts the Cox-Snell R-squared to be used with the expected outcome prevalence (0.28). A global shrinkage factor of 0.90 was also considered, totaling a minimum sample size of 1657 individuals with at least 464 events for model fitting.

For the CSL dataset, a maximum Cox–Snell R2 statistic of 0.1 and a global shrinkage factor of 0.90 were employed, considering an overall prevalence of 0.16, which totaled 3651 individuals with at least 585 events as a minimum sample size for model fitting.[27]

The primary outcome was delivery mode, with VD corresponding to 0 and CS corresponding to 1. To simplify and develop an algorithm with binary output, the VD category includes instrumental deliveries.

Model development

Models were developed on the Portuguese dataset and internally validated using 10-fold stratified cross-validation, which divides the dataset into 10 parts, 9 for training and 1 for testing, with subsequent 30 runs with different random initializations.[10] Multiple ML methods were tested and compared, namely, logistic regression (LR), multi-layer perceptron, random forest, support vector classifier, extreme gradient boosted trees (XGBoost), and AdaBoost classifiers.[28] All models were built using the scikit-learn (https://scikit-learn.org/) and XGBoost (https://xgboost.readthedocs.io/en/stable/) implementations with the default off-the-shelf settings.

The best model´s 43 feature importance evaluations were made by running a SHapley Additive exPlanation (SHAP) feature importance analysis.[29] With features ranked by importance, the elbow method was applied to find the minimum number of features required to perform well without

compromising predictive power. This method incrementally adds new features until performance stagnates. This evaluation was conducted by retraining all these variations using 10-fold cross-validation of the Portuguese dataset and analyzing their performance metrics. Then, the best-performing model was retrained with the entire Portuguese dataset and tested over the whole CSL cohort for external validation.[8]

Performance metrics (F1-score, sensitivity, specificity, positive predictive value (PPV), and negative predictive value) and respective confusion matrices (CM) were assessed separately for each internal and external validation dataset. The most relevant metrics for model selection were F1-score and PPV because we aim to avoid performing CS in women who would have VD (false positive - FP) and correctly identify true CS (true positive) while considering the class imbalance of both datasets.[30]

Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis – Artificial Intelligence (TRIPOD-AI) recommendations were followed on the visual analysis of model scores: (1) discrimination, using the area under the receiver operating characteristic (AUROC) curve, (2) calibration, and (3) decision curve analysis (DCA).[31] AUROC measures the sensitivity and specificity of a model for different cut-offs, quantifying the model´s ability to discriminate between outcomes.[7] Calibration measures the proportion of individuals with a risk prediction of actually experiencing the outcome, visualized by plotting the predicted outcome against the observed outcome probability.[32] DCA assesses the benefits (performing a CS on a true positive case) and harms (performing a CS on an FP case) associated with the model.[17] A comparison was made between the default strategies, treating all (performing a CS) and treating none (all VD), for both dataset models.[33] The mean evaluation metric values for each k-fold of the internal validation models were reported with 95% CIs.

Statistical analysis

Experimental and statistical analyses were performed using Python. Descriptive statistics were used to present categorical numerical features. Student-t test was used for continuous parametric features; Mann-Whitney U test for non-parametric continuous features, and Chi-square test for categorical features. Data normality and equality of variance were evaluated using Kolmogorov-Smirnov and Levene's tests. The Wilcoxon test assessed the difference between the best model and the others. Statistical significance was set at $p<0.05$.

Ethics statement

The Ethics Committee from Local Health Unit of Coimbra reviewed and approved the study on 22 April 2022 (CE-047/2022). Requirement for written informed consent was waived due to the retrospective, de-identified nature of the patient data. We adhered to the TRIPOD-AI statement (Supplementary Data —TRIPOD Checklist).

DECLARATION STATEMENTS

Data availability

Data cannot be made publicly available due to ethical approval and contractual restrictions.

Code availability

We share our code and associated information on the following GitHub link: https://github.com/PugtgYosuky/AI4Birth.

Abbreviations: Consortium on Safe Labor (CSL), confusion matrix (CM), cesarean section (CS), decision curve analysis (DCA), false positive – FP, labor induction (IOL), logistic regression (LR), machine learning (ML), positive predictive value (PPV), SHapley Additive exPlanation (SHAP), under the receiver operating characteristic (AUROC), Vaginal delivery (VD).

Acknowledgments

The data included in this paper were obtained from the Consortium on Safe Labor, supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, through Contract No. HHSN267200603425C. Institutions involved in the Consortium include, in alphabetical order: Baystate Medical Center, Springfield, MA; Cedars-Sinai Medical Center Burnes Allen Research Center, Los Angeles, CA; Christiana Care Health System, Newark, DE; Georgetown University Hospital, MedStar Health, Washington, DC; Indiana University Clarian Health, Indianapolis, IN; Intermountain Healthcare and the University of Utah, Salt Lake City, Utah; Maimonides Medical Center, Brooklyn, NY; MetroHealth Medical Center, Cleveland, OH.; Summa Health System, Akron City Hospital, Akron, OH; The EMMES Corporation, Rockville MD (Data Coordinating Center); University of Illinois at Chicago, Chicago, IL; University of Miami, Miami, FL; and University of Texas Health Science Center at Houston, Houston, Texas. The named authors alone

are responsible for the views expressed in this manuscript, which does not necessarily represent the decisions or the stated policy of the NICHD.

We acknowledge NICHD DASH for providing the Consortium on Safe Labor data that was used for this research.

This work has received funding from the Research Grant from the Coimbra Hospital and University Centre.

Author Contributions:

Iolanda Ferreira: Conceptualization, Methodology, Validation, Investigation, Data Curation, Writing - Original Draft, Funding acquisition;

Joana Simões: Methodology, Software, Formal analysis, Resources, Data Curation

João Nuno Correia: Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Supervision;

Ana Luísa Areia: Conceptualization, Methodology, Validation, Writing - Review & Editing, Supervision;

Competing interests:

The authors declare no competing interests.

References

1 Middleton, P., Shepherd, E., Morris, J., Crowther, C. A. & Gomersall, J. C. Induction of labour at or beyond 37 weeks' gestation. *Cochrane Database Syst Rev* **7**, CD004945, doi:10.1002/14651858.CD004945.pub5 (2020).

2 Grobman, W. A. & Caughey, A. B. Elective induction of labor at 39 weeks compared with expectant management: a meta-analysis of cohort studies. *Am J Obstet Gynecol* **221**, 304-310, doi:10.1016/j.ajog.2019.02.046 (2019).

3 Kawakita, T., Iqbal, S. N., Huang, C. C. & Reddy, U. M. Nonmedically indicated induction in morbidly obese women is not associated with an increased risk of cesarean delivery. *Am J Obstet Gynecol* **217**, 451 e451-451 e458, doi:10.1016/j.ajog.2017.05.048 (2017).

4 Rossi, R. M. *et al.* Risk Calculator to Predict Cesarean Delivery Among Women Undergoing Induction of Labor. *Obstet Gynecol* **135**, 559-568, doi:10.1097/AOG.0000000000003696 (2020).

5 American College of, O. a. & Gynecologists, C. A. G., Raghuraman N., Gandhi M., Kaimal A.J. Clinical Practice Guideline No. 8, First and Second Stage Labor Management. *Obstet Gynecol* **143**, 144-162, doi:10.1097/AOG.0000000000005447 (2024).

6 Hamilton, E. F. *et al.* New labor curves of dilation and station to improve the accuracy of predicting labor progress. *Am J Obstet Gynecol* **231**, 1-18, doi:10.1016/j.ajog.2024.02.289 (2024).

7 Levine, L. D. *et al.* A validated calculator to estimate risk of cesarean after an induction of labor with an unfavorable cervix. *Am J Obstet Gynecol* **218**, 254 e251-254 e257, doi:10.1016/j.ajog.2017.11.603 (2018).

8 Tsur, A. *et al.* Development and validation of a machine-learning model for prediction of shoulder dystocia. *Ultrasound Obstet Gynecol* **56**, 588-596, doi:10.1002/uog.21878 (2020).

9 Meyer, R. *et al.* Utilizing machine learning to predict unplanned cesarean delivery. *Int J Gynaecol Obstet* **161**, 255-263, doi:10.1002/ijgo.14433 (2023).

10 D'Souza, R. *et al.* Prediction of successful labor induction in persons with a low Bishop score using machine learning: Secondary analysis of two randomized controlled trials. *Birth* **50**, 234-243, doi:10.1111/birt.12691 (2023).

11 Hu, T. *et al.* Establishment of a model for predicting the outcome of induced labor in full-term pregnancy based on machine learning algorithm. *Sci Rep* **12**, 19063, doi:10.1038/s41598-022-21954-2 (2022).

12 Ferreira, I., Simoes, J., Correia, J. & Areia, A. L. Predicting vaginal delivery after labor induction using machine learning: Development of a multivariable prediction model. *Acta Obstet Gynecol Scand*, doi:10.1111/aogs.14953 (2024).

13 Jochum, F. *et al.* Externally Validated Score to Predict Cesarean Delivery After Labor Induction With Cervi Ripening. *Obstet Gynecol* **134**, 502-510, doi:10.1097/AOG.0000000000003405 (2019).

14 Steinberg, S., Wong, M., Zimlichman, E. & Tsur, A. Novel Machine Learning Applications in Peripartum Care: A Scoping Review. *Am J Obstet Gynecol MFM*, 101612, doi:10.1016/j.ajogmf.2025.101612 (2025).

15 Kawakita, T. *et al.* Externally Validated Prediction Model of Vaginal Delivery After Preterm Induction With Unfavorable Cervix. *Obstet Gynecol* **136**, 716-724, doi:10.1097/AOG.0000000000004039 (2020).

16 Shao, S. J., Teal, E. N., Lewkowitz, A. K., Gaw, S. L. & Sobhani, N. C. Validated Calculators Predicting Cesarean Delivery After Induction: Accuracy in an External Population. *Obstet Gynecol* **142**, 99-107, doi:10.1097/AOG.0000000000005234 (2023).

17 Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* **14**, 49-58, doi:10.1093/ckj/sfaa188 (2021).

18 Zhang, J. *et al.* Contemporary patterns of spontaneous labor with normal neonatal outcomes. *Obstet Gynecol* **116**, 1281-1287, doi:10.1097/AOG.0b013e3181fdef6e (2010).

19 Kawakita, T., Saeed, H. & Huang, J. C. An Externally Validated Model to Predict Prolonged Induction of Labor with an Unfavorable Cervix. *Am J Perinatol* **41**, e3140-e3146, doi:10.1055/a-2195-6063 (2024).

20 Moons, K. G. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* **162**, W1-73, doi:10.7326/M14-0698 (2015).

21 Bushman, E. T. *et al.* Influence of Estimated Fetal Weight on Labor Management. *Am J Perinatol* **37**, 252-257, doi:10.1055/s-0039-1695011 (2020).

22 de Hond, A. A. H. *et al.* Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med* **5**, 2, doi:10.1038/s41746-021-00549-7 (2022).

23 Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). The Consortium on Safe Labor. Available at: https://www.nichd.nih.gov/research/supported/safe-labor (Accessed: 2025-08-28).
24 #14. National Institute for Health and Care Excellence (NICE). Inducing Labour. NICE Guideline [NG207]. 2021. Available at: https://www.nice.org.uk/guidance/ng207/resources/inducing-labour-pdf-66143719773637
25 American College of, O. a. & Gynecologists, R. M., Ramin S., ACOG Committe on Practice Bulletins. Practice Bulletin No. 107: Induction of labor. *Obstet Gynecol* **114**, 386-397, doi:10.1097/AOG.0b013e3181b48ef5 (2009).
26 Mack, C., Su, Z. & Westreich, D. in *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition AHRQ Methods for Effective Health Care* (2018).
27 Riley, R. D. *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* **368**, m441, doi:10.1136/bmj.m441 (2020).
28 Lipschuetz, M. *et al.* Prediction of vaginal birth after cesarean deliveries using machine learning. *Am J Obstet Gynecol* **222**, 613 e611-613 e612, doi:10.1016/j.ajog.2019.12.267 (2020).
29 Lundberg SM, S.-I. L. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
30 Ferreira, I., Simoes, J., Pereira, B., Correia, J. & Areia, A. L. Ensemble learning for fetal ultrasound and maternal-fetal data to predict mode of delivery after labor induction. *Sci Rep* **14**, 15275, doi:10.1038/s41598-024-65394-6 (2024).
31 Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, e078378, doi:10.1136/bmj-2023-078378 (2024).
32 Riley, R. D. *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* **40**, 4230-4251, doi:10.1002/sim.9025 (2021).
33 Kyriacou, C. *et al.* Updating M6 pregnancy of unknown location risk-prediction model including evaluation of clinical factors. *Ultrasound Obstet Gynecol* **63**, 408-418, doi:10.1002/uog.27515 (2024).

Tables

Table 1: Internal and External Validation Cohort Characteristics

Demographic data	**Internal validation dataset (N=2434)**	**External validation dataset (N= 10591)**	**p-value**
Maternal age in years (mean, SD)	32.7 (5.5)	28.1 (5.4)	<**0.001**
Maternal height in cm (mean, SD)	163.0 (6.0)	165.0 (7.4)	<**0.001**
Nulliparity (%)	1417 (58.2%)	4120 (38.9%)	<**0.001**
Previous CS (%)	275 (11.3%)	213 (2.0%)	<**0.001**
Pre-pregnancy body mass index (kg/m^2) (mean, SD)	25.6 (5.2)	25.8 (6.0)	0.29
Final pregnancy weight (kg) (mean, SD)	81.6 (14.7)	85.5 (17.3)	<**0.001**
IOL data			
IOL method (%)			<**0.001**
Misoprostol	996 (40.9%)	1493 (14.1%)	<**0.001**
Dinoprostone	1325 (54.4%)	1280 (12.1 %)	<**0.001**

Foley catheter	58 (2.4%)	1176 (11.0%)	**0.004**
Oxytocin	36 (1.5%)	7712 (72.8%)	**<0.001**
Gestational age at IOL (Mean, SD)	39.8 (1.2)	39.1 (1.5)	**<0.001**
Indication for labor induction			
Post-date pregnancy (%)	1046 (43.0%)	1457 (13.8%)	**<0.001**
Fetal complications (%)	262 (10.8%)	1205 (11.4%)	0.41
Maternal pregnancy complications (%)	390 (16.0%)	723 (6.8%)	**<0.001**
Hypertensive disorders of pregnancy (%)	176 (7.2%)	1217 (11.5%)	**<0.001**
Maternal previous pathologies (%)	51 (2.1%)	18 (0.2%)	**<0.001**
Premature rupture of membranes (%)	143 (5.9%)	239 (2.3%)	**<0.001**
Fetal pathology (%)	29 (1.2%)	74 (0.7%)	0.02
Other (%)	337 (13.8%)	5806 (54.8%)	**<0.001**
BISHOP			
0-3 (%)	1968 (80.9%)	1838 (17.4%)	**<0.001**
4-5 (%)	430 (17.7%)	3170 (29.9%)	0.09
6-7 (%)	36 (1.5%)	5583 (52.7%)	0.72
Delivery data			
Delivery mode (%)			
Eutocic	1009 (41.4%)	8021 (75.7%)	**<0.001**
Vaginal instrumental	727 (29.9%)	834 (7.9%)	**<0.001**
Cesarean section	698 (28.7%)	1736 (16.4%)	**<0.001**
Cesarean section indications (%)			
FHR abnormality	214 (8.0%)	477 (4.5%)	<0.001
Failed induction	187 (7.7%)	10 (0.1%)	<0.001
Labor dystocia	274 (11.3%)	38 (0.4%)	<0.001
Other	23 (0.9%)	1211 (11.4%)	<0.001
Post-partum and Neonatal data			
Neonatal birthweight (grams, SD)	3319.0 (459.5)	3372.0 (500.4)	**<0.001**
5-minute Apgar Score ≤ 7 (%)	19 (0.8%)	240 (2.3%)	**<0.001**
Neonatal intensive care admission (%)	42 (1.7%)	912(8.6 %)	**<0.001**
Post-partum hemorrhage (%)	97 (4.0%)	324 (3.1%)	0.02
Third and fourth degree perineal tear (%)	7 (0.3%)	242 (2.3%)	**<0.001**

CS: cesarean section; FHR: fetal heart rate; IOL: induction of labor; kg: kilograms; SD: standard deviation.

Table 2: Performance metrics of model development and evaluation

Model	AUROC	F1-score	PPV	NPV	Sensitivity	Specificity	p-value
Portuguese dataset - CV - All Features							
Logistic Regression	0.793 (0.030)	0.748 (0.023)	0.752 (0.026)	0.793 (0.015)	0.767 (0.021)	0.911 (0.023)	Reference
AdaBoost Classifier	0.772 (0.031)	0.739 (0.024)	0.739 (0.027)	0.791 (0.016)	0.755 (0.023)	0.893 (0.025)	**< 0.001**
SVC	0.760 (0.031)	0.726 (0.023)	0.747 (0.029)	0.772 (0.013)	0.760 (0.019)	0.942 (0.019)	**< 0.001**
Random Forest Classifier	0.759 (0.032)	0.722 (0.024)	0.728 (0.029)	0.775 (0.014)	0.748 (0.022)	0.912 (0.023)	**< 0.001**
MLP Classifier	0.753 (0.033)	0.734 (0.025)	0.731 (0.027)	0.795 (0.018)	0.745 (0.025)	0.865 (0.028)	**< 0.001**
XGB Classifier	0.748 (0.032)	0.719 (0.025)	0.716 (0.027)	0.785 (0.017)	0.731 (0.024)	0.859 (0.027)	**< 0.001**
Portuguese dataset - CV - Reduced Features							
LR - 13 Best Features	0.799 (0.029)	0.750 (0.023)	0.756 (0.025)	0.794 (0.015)	0.770 (0.020)	0.916 (0.022)	**0.015**
Train Portuguese Dataset - Test CSL Dataset							
LR - All Features	0.813	0.788	0.830	0.846	0.845	0.995	-
LR - 13 Best Features	0.808	0.781	0.822	0.843	0.842	0.995	-

Results are presented as mean (standard deviation) when addressing the Portuguese dataset. These are the average results obtained after 30 runs with different initializations. Each run yields an average value obtained by 10-fold cross-validation. Mean and standard deviation values are not presented when the LR model is tested on the CSL dataset, because we are testing on the entire dataset. Also, there are no differences in values because Sklearn's default LR has no stochastic component. AUROC: area under the receiver operating curve; CV: cross-validation; CSL: Consortium for Safe Labor; LR: Logistic regression; MLP: multi-layer perceptron; NPV: negative predictive value; PPV: positive predictive value; SVC: support vector classifier; XGB: extreme gradient boosted trees; p-value was obtained by the Wilcoxon test comparing model distribution with reference model, Logistic Regression across all reported metrics..

Figure Legends

Figure 1: SHAP analysis of the most influential features on the logistic regression model in the Portuguese dataset. The SHAP plots were cross-validated and represent the mean of each fold. The SHAP plot shows the effect of each feature on the model's prediction score, by order of importance. On the x-axis, higher values appear more red and lower values more blue. Each point represents an individual case. If the dots are increasingly red or blue on one side of the central line, increasing or decreasing values move the mode prediction in that direction. The x-axis corresponds to the impact on log-odds of each variable. CS: cesarean section; SHAP: SHapley Additive exPlanation; VD: vaginal delivery.

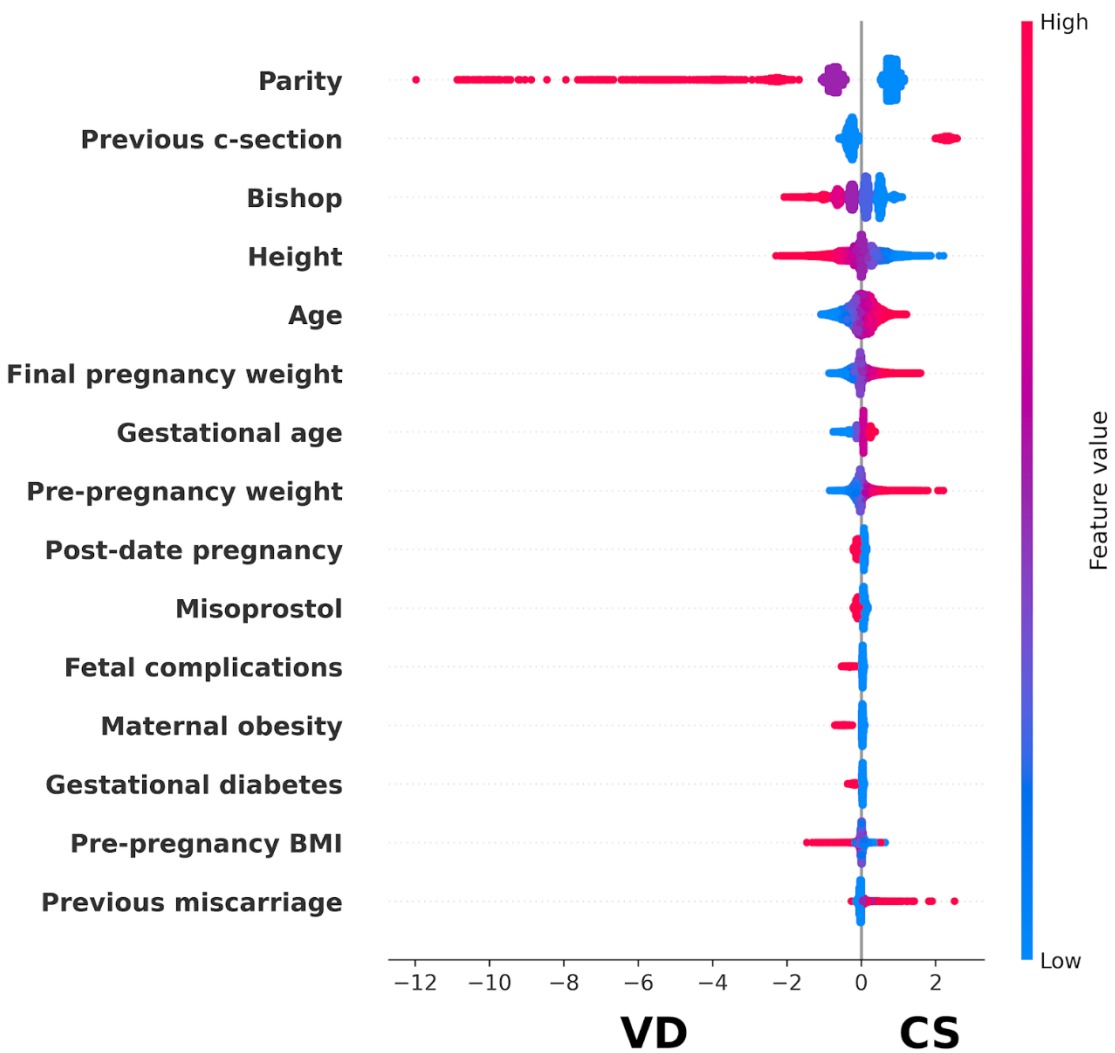

Figure 2: AUROC curves of the LR model with all features and the simplified LR model, first in the Portuguese dataset using cross-validation (left) and second, training with the Portuguese dataset and testing the CSL dataset (right). CSL: Consortium for Safe Labor; LR: Logistic regression; PT: Portuguese.

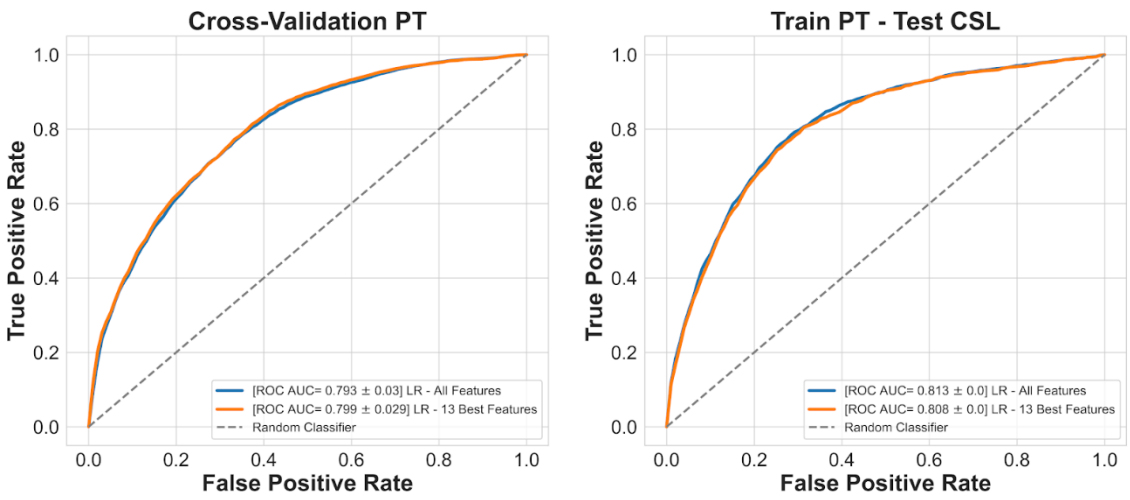

Figure 3: Calibration curves and respective intercept (C-I) and slope (C-S) scores for the Portuguese dataset using cross-validation (left) and training with the Portuguese dataset and testing the CSL dataset (right). The dashed line indicates perfect agreement between the predicted probability of the model and the actual probability. CSL: Consortium for Safe Labor; LR: Logistic Regression; PT: Portuguese.