The Journal of Maternal-Fetal & Neonatal Medicine

ISSN: 1476-7058 (Print) 1476-4954 (Online) Journal homepage: http://www.tandfonline.com/loi/ijmf20

Risk of operative delivery for intrapartum fetal compromise in small-for-gestational-age fetuses at term: external validation of the IRIS algorithm

Erkan Kalafat, Jose Morales-Rosello, Elisa Scarinci, Basky Thilaganathan & Asma Khalil

To cite this article: Erkan Kalafat, Jose Morales-Rosello, Elisa Scarinci, Basky Thilaganathan & Asma Khalil (2018): Risk of operative delivery for intrapartum fetal compromise in small-for-gestational-age fetuses at term: external validation of the IRIS algorithm, The Journal of Maternal-Fetal & Neonatal Medicine, DOI: 10.1080/14767058.2018.1560412

To link to this article: https://doi.org/10.1080/14767058.2018.1560412

Accepted author version posted online: 18 Dec 2018.

Risk of operative delivery for intrapartum fetal compromise in small-for-gestational-age fetuses at term: external validation of the IRIS algorithm

Erkan Kalafat, MD[*†‡]

Jose Morales-Rosello, MD[¶]

Ms. Elisa Scarinci[¶]

Basky Thilaganathan, PhD, FRCOG[*§]

Asma Khalil, MRCOG [*§]

*Fetal Medicine Unit, St George's University Hospitals NHS Foundation Trust, London, UK

†Ankara University Faculty of Medicine, Department of Obstetrics and Gynecology, Ankara, TURKEY

‡Middle East Technical University, Department of Statistics, Ankara, TURKEY

¶Hospital Universitario y Politecnico la Fe, Department of Obstetrics and Gynecology, Valencia, SPAIN

§Molecular and Clinical Sciences Research Institute, St. George's University of London, London, UK

Funding details: None

Disclosure statement: Nothing to disclose.

Corresponding Author

Professor Asma Khalil

Fetal Medicine Unit

St George's University of London

London SW17 0RE

Telephone: (Work) +442032998256

Mobile: +447917400164.

Fax: +442077339534

E-mail: akhalil@sgul.ac.uk; asmakhalil79@googlemail.com"

Short title: Prediction of fetal compromise

KEY WORDS

SGA, adverse outcome, prediction, risk assessment, Doppler, CPR, mobile app

ABSTRACT

OBJECTIVES

Small-for-gestational-age fetuses (SGA) are at high-risk of intrapartum fetal compromise requiring operative delivery. In a recent study, we developed a model using a combination of three antenatal (gestational age at delivery, parity, cerebroplacental ratio) and three intrapartum (epidural use, labor induction and augmentation using oxytocin) variables for the prediction of operative delivery due to presumed fetal compromise in SGA fetuses – the **I**ndividual **RI**sk a**S**sessment (IRIS) prediction model. The aim of this study was to test the predictive accuracy of the IRIS prediction model in an external cohort of singleton pregnancies complicated by SGA.

METHODS

This was an external validation study using a cohort of pregnancies from two tertiary referral centers in Spain and England. The inclusion criteria were singleton pregnancies diagnosed with an SGA fetus, defined as EFW below the 10th centile for gestational age at 36 weeks or beyond, which had fetal Doppler assessment and available data on their intrapartum care and pregnancy outcomes. The main outcome in this study was the operative delivery for presumed fetal compromise. External validation was performed using the coefficients obtained in the original development cohort. The predictive accuracies of models were investigated with receiver operating characteristics (ROC) curves. The Hosmer-Lemeshow test was used to test the goodness-of-fit of models and calibration plots were also obtained for visual assessment. A mobile application using the combined model algorithm was developed to facilitate clinical use.

RESULTS

412 singleton pregnancies with an antenatal diagnosis of SGA were included in the study. The operative delivery rate was 22.8% (n=94). The group which

required operative delivery for presumed fetal compromise had significantly fewer multiparous women (19.1% vs 47.8%, p<0.001 in the total study population; 19.0% vs 43.5% and 19.2% vs. 49.6%, UK and Spain cohort, respectively), lower CPR MoM (median: 0.77 vs 0.92, p<0.001 in the total study population; 0.77 vs 0.92 and 0.77 vs 0.92 , UK and Spain cohort, respectively), more inductions of labor (74.5% vs 60.1%, p=0.010 in the total study population; 85.7% vs 77.2% and 71.2% and 53.1, UK and Spain cohort, respectively) and more use of oxytocin augmentation (57.4% vs 39.3%, p=0.002 in the total study population; 19.0% vs 12.0% and 68.5% and 50.4%, UK and Spain cohort, respectively) compared to those who did not require operative delivery due to presumed fetal compromise. When the original antenatal model was applied to the present cohort, we observed moderate predictive accuracy (AUC: 0.70, 95% CI: 0.64-0.76), and no signs of poor fit (p=0.464). The original combined model, when applied to the external cohort, had moderate predictive accuracy (AUC: 0.72, 95% CI: 0.67-0.77) and also no signs of poor fit (p=0.268) without the need for refitting. A statistically significant increase in the predictive accuracy was not achieved via refitting of the combined model (AUC 0.76 vs 0.72, p=0.060).

CONCLUSIONS

Using our recently published model, the predictive accuracy for fetal compromise requiring operative delivery in term fetuses thought to be SGA was modest and showed no signs of poor fit in an external cohort. The IRIS tool for mobile devices has been developed to facilitate wide clinical use of this prediction model.

Brief rationale

Objective: To determine the external validity of an intrapartum risk prediction model for suspected small-for-gestational age fetuses.

What is already known: Small-for-gestational age foetuses are at increased risk of intrapartum compromise. Fetal weight alone is a poor marker for adverse outcomes and a comprehensive prediction model has been previously suggested.

What this study adds: Multivariable prediction model showed good accuracy and calibration in this external validation study. The significance of some variables was different between the original and external validation cohort and there was a small margin for improvement with model refitting. A mobile application has been developed to facilitate clinical use.

INTRODUCTION

Small-for-gestational-age fetuses (SGA) are at high-risk of intrapartum fetal compromise, operative delivery, perinatal morbidity and demise [1-6]. Failure to detect SGA fetuses during the antenatal period is associated with increased perinatal morbidity and mortality [7-9]. However, the incidence of adverse outcome in the pregnancies with SGA fetus near term is relatively small [10]. Differentiation of cases truly at risk for adverse outcomes from constitutionally small fetuses is therefore essential for the management of SGA. Unfortunately, estimated fetal weight (EFW) alone is a poor predictor of adverse outcomes, and therefore, additional markers are needed [11-13]. Cerebroplacental ratio (CPR) has emerged as an important Doppler index for the prediction of adverse outcomes in SGA fetuses [14-17]. A low CPR is associated with increased risk of neonatal unit admission, intrapartum fetal compromise and need for operative delivery [14-17]. When used in conjunction with other antenatal and intrapartum variables, CPR has the potential to be used for the prediction of adverse outcomes such as operative delivery and neonatal unit admission [4,6].

In a recent study, we developed a model using a combination of three antenatal and three intrapartum variables for the prediction of operative delivery due to presumed fetal compromise in SGA fetuses – the **I**ndividual **RI**sk a**S**sessment (IRIS) prediction model [4]. Such a prediction model could be helpful for risk stratification of SGA fetuses and patient counseling regarding the timing and mode of birth. However, prediction models are known to overestimate the predictive accuracy in the development cohort. Validation studies are required to assess the performance of such models in different populations [18]. The aim of this study was to test the predictive accuracy of the IRIS prediction model in an external cohort of SGA pregnancies.

METHODS

This was an external validation study using two cohorts of pregnancies from two tertiary referral centers in Spain and England. Pregnancies over a 5-year period (2012-2017) in Spain (Hospital Universitario y Politecnico la Fe) and over a 2-year period (2016-2018) in UK (St. George's University Hospital) were used. None of these pregnancies were included in the development of the prediction model. The inclusion criteria were singleton pregnancies diagnosed with an SGA fetus, defined as EFW below the 10th centile for gestational age at 36 weeks or beyond with complete Doppler assessment (including umbilical and middle cerebral artery) within one month prior to the delivery and also complete intrapartum information (induction, augmentation and epidural use). Further details on how the prediction models were built can be found in the original study. [4] Model variables were routinely recorded in St. George's Hospital (UK cohort). In the Universitario y Politecnico la Fe Hospital (Spanish cohort), Doppler variables were not routinely recorded, and therefore only the cases with complete fetal Doppler assessment (which were routinely performed by a single operator; JM) were included in the analysis. Pregnancies complicated by major structural fetal abnormalities, aneuploidy or genetic syndromes were excluded from the analysis. Moreover, pregnancies that had an elective cesarean delivery were also excluded from the analysis. Gestational age (GA) was calculated from the crown-rump length measurement at 11-13 weeks and only one (the last) examination per pregnancy was included in the analysis [19]. Rarely, for pregnancies in which the first ultrasound was performed in the second trimester (>14 weeks' gestation), the pregnancy was dated according to the head circumference. Routine fetal biometry was performed according to a standard protocol and the EFW was calculated using the Hadlock formula [20].The umbilical artery (UA) and middle cerebral artery (MCA) Doppler waveforms were recorded using color Doppler, and the pulsatility index (PI) was calculated according to a standard protocol [21]. In brief, the MCA PI values were obtained in the space where the artery passes by the

sphenoid wing close to the Circle of Willis, and the UA PI values were obtained in free loops of umbilical cord. The Doppler measurements were performed within four weeks of delivery. The measurements were obtained in the absence of fetal movement, and keeping the insonation angle with the examined vessels less than 30°.The CPR was calculated as the simple ratio between the MCA PI and the UA PI. All Doppler indices and biometry variables were converted into multiples of median (MoM) and centiles correcting for GA using reference ranges, and birthweight values were converted into centiles [22-24]. Antenatal follow-up and delivery were managed by separate teams. The CPR and relevant measures (multiples of median) were not calculated before the analysis of this study.

Intrapartum data included whether the labor was induced or spontaneous, use of oxytocin for slow progress of labor, use of epidural analgesia for labor, and mode of delivery. Data on the maternal baseline characteristics and the pregnancy outcomes were collected from the hospital obstetric records. The main outcome in this study was the operative delivery for presumed fetal compromise. Operative delivery for presumed fetal compromise included both cesarean delivery and instrumental delivery. The diagnosis of fetal compromise was based on CTG abnormalities (as defined in Sociedad Espanola de Ginecologia y Obstetricia (SEGO) and National Institute for Health and Care Excellence (NICE) guidelines for Spain and UK cohorts, respectively), abnormal fetal scalp blood sample pH (pH<7.20), or a combination of these [25,26].SEGO and NICE guidelines are similar to one other in terms of criteria for diagnosing abnormal CTG traces.

Statistical analysis

Continuous variables were presented as median with interquartile range, while binary variables were presented as a fraction of the total with percentages.

Distribution assumptions were tested with Shapiro-Wilk test. Group comparisons of variables were performed with t-test, Mann-Whitney-U test or Fisher's exact test where appropriate. External validation was performed using the coefficients obtained in the original development cohort. Possible explanations for poor fitting or suboptimal predictive accuracy were investigated by refitting the model to the external validation cohort. The cohort site (England or Spain) was used as a factor and possible interactions between explanatory variables and cohort site was investigated during model refitting ($P<0.10$ was deemed significant for interaction). Parameter estimates of the original model and refitted models were used to predict the probability of operative delivery in the external validation cohort using an inverse logit function. The predicted probabilities were then used for diagnostic procedures, such as predictive accuracy and goodness-of-fit assessment. The predictive accuracies of models were investigated with receiver operating characteristics (ROC) curves. The accuracy values (true positive + true negative / total) of each model for different risk cut-offs were also calculated. A Bayesian framework was used to calculate the posterior probability of improvement in accuracy, sensitivity and specificity for each risk cut-off. The predictive markers of accuracy, sensitivity and specificity were modeled using binomial distribution for the likelihood function and Beta distribution (0.5,0.5) for the prior function. A random-walk metropolis algorithm was used and Markov Chain Monte Carlo (MCMC) simulations were run for 200.000 iterations. Separate ROC curves were also obtained for two cohort sites to investigate the site-specific differences in the predictive accuracy of the model. Comparisons of ROC curves were made with De Long's test. The confidence intervals for the accuracy values were calculated with 10.000 stratified bootstrap replicates. The Hosmer-Lemeshow test was used to test the goodness-of-fit of models and calibration plots were also obtained for visual assessment. We aimed for approximately 100 operative delivery cases in our validation cohort as suggested in the literature for minimizing optimism in external validation studies [27]. The statistical analysis

was performed using the RStudio (Version 1.1.419, RStudio, Inc.) statistical software [28].

The **I**ndividual **RI**sk a**S**sessment (IRIS) mobile application

The algorithm used in this study was implemented in a mobile phone application and it is available free-of-charge for Android and iOS mobile devices in their respective application stores (https://goo.gl/qo31Rm & https://itunes.apple.com/us/app/iris-tool-for-sga-babies/id1371991518?ls=1&mt=8). The mobile application uses the combined model and requires gestational age at delivery, parity, CPR, epidural use, labor augmentation and induction information to calculate individual risk of operative delivery in terms of percentage probability.

RESULTS

We identified 490 singleton pregnancies with an antenatal diagnosis of SGA fetus which were eligible for inclusion (344 in the Spanish cohort and 146 in the UK cohort). After excluding elective Cesarean sections (n=66), fetal anomalies (n=3), stillbirth cases (n=3), and cases with missing variables (n=6), 412 pregnancies with an operative delivery rate of 22.8% (n=94) were included in the study (Figure 1). The total number of operative delivery cases due to presumed fetal compromise was 73 (24.4%) in the Spanish cohort and 21 (18.6%) in the UK cohort. The incidence of Cesarean delivery rate in the study cohort was 14.3% (n=59) and that of the instrumental delivery due to presumed fetal compromise was 8.5% (n=35). The accuracy of the antenatal ultrasound to detect SGA at birth was 87.6% and 82.6% for UK and Spain cohort, respectively.

The antenatal, intrapartum and birth characteristics of the study cohort stratified according to the location are shown in Table 1. The group which required operative delivery for presumed fetal compromise had significantly fewer multiparous women (19.1% vs 47.8%, p<0.001 in the total study population; 19.0% vs 43.5% and 19.2% vs. 49.6%, UK and Spain cohort, respectively), lower CPR MoM (median: 0.77 vs 0.92, p<0.001 in the total study population; 0.77 vs 0.92 and 0.77 vs 0.92 , UK and Spain cohort, respectively), more inductions of labor (74.5% vs 60.1%, p=0.010 in the total study population; 85.7% vs 77.2% and 71.2% and 53.1, UK and Spain cohort, respectively) and more use of oxytocin augmentation (57.4% vs 39.3%, p=0.020 in the total study population; 19.0% vs 12.0% and 68.5% and 50.4%, UK and Spain cohort, respectively) compared to those who did not require operative delivery due to presumed fetal compromise. The neonates delivered via operative delivery for presumed fetal compromise also had lower birthweight centiles (P=0.014 and P<0.001, UK and Spain cohort, respectively).

The validation of the original antenatal and combined models was performed using the original model coefficients and with re-estimated model coefficients (Table 2). When the original antenatal model was applied to the present cohort, we observed moderate predictive accuracy (AUC 0.70, 95% CI 0.64-0.76, Figure

2), and good fit (p=0.464). Visual estimation of the model fit with calibration plot showed that the original antenatal model slightly underestimated the probability of operative delivery due to presumed fetal compromise (Supplementary Figure 1). When the variables were refitted to the present cohort, only parity and CPR MoM remained as significant predictors (Table 2). The refitted model showed similar predictive accuracy compared to the original antenatal model (AUC 0.73, 95% CI 0.67-0.79; De Long's test, p=0.076, Figure 1) and poor-fit according to Hosmer-Lemeshow test (p=0.008) (Supplementary Figure 2).

When the original combined model was applied to the present cohort, we observed moderate predictive accuracy (AUC 0.72, 95% CI 0.67-0.77, Figure 3) and good fit (p=0.268). When variables were refitted to the present cohort, only parity, CPR MoM, labor induction and oxytocin augmentation remained significant predictors (Table 2). Furthermore, when the cohort site was used as a factor, there was a borderline significant effect for interaction between cohort site and epidural use (p=0.079) and cohort site was a borderline significant factor on its own as well (p=0.075). The refitted model showed statistically non-significant improvement in the predictive accuracy compared to the original antenatal model (AUC 0.76, 95% CI 0.70-0.81 vs AUC 0.72, 95% CI 0.67-0.77, p=0.060, Figure 3) and also good-fit (p=0.545) (Supplementary Figures 3 and 4). The predictive accuracy values of the original and refitted combined models can be seen in Table 3. The risk cut-offs between 30% and 40% offered the most balanced sensitivity and specificity values. The posterior probabilities indicated that the chance of accuracy improvement with model refitting is highly probable for risk cut-off ranges between 10% and 40% at the cost of reduced sensitivity (Table 3). We also performed a sensitivity analysis for each cohort site and found no evidence of significant differences in the predictive accuracy according to ROC curves (Supplementary Figure 5).

The external validation shows that the original combined model has moderate predictive accuracy (AUC 0.72) and goodness-of-fit (p=0.268) when used in an external cohort without the need for refitting.

DISCUSSION

Main findings

The predictive accuracy for fetal compromise requiring operative delivery of our model was modest and showed no signs of poor fit in an external cohort of pregnancies with suspected SGA fetuses at term. Although the re-estimation of combined model variables did not significantly improve the predictive accuracy Bayesian framework analysis indicated accuracy improvement is probable at the expense of sensitivity. The IRIS application for mobile devices has been developed to facilitate the clinical use of this prediction model.

Interpretation of the findings and comparison with existing literature

There were some demographic and clinical differences between the original cohort and the external validation cohort. When the model parameters were re-estimated using the validation cohort, we observed some variables - namely epidural use and GA at delivery - which no longer appeared important for predicting operative delivery. The estimated effect of GA at delivery was similar to the original cohort, albeit with a larger confidence interval probably because of the smaller number of pregnancies in the external validation cohort compared to the original cohort. This finding may affect the 95% CI estimation, but does not necessarily imply a lack of casual relationship. Interestingly, the direction of effect of the epidural analgesia was towards reduced risk in the external validation cohort, which is contrary to the original cohort. However, there was a borderline significant interaction between the cohort site and epidural use with the UK cohort showing an increased risk with the epidural use (P=0.079). This difference can be explained by the effect of epidural analgesia on labor outcomes, which is quite heterogeneous in the literature and is influenced by local clinical practices [29]. Furthermore, the original and part of the validation cohort used different guidelines for the diagnosis of presumed fetal compromise that could also explain the observed difference in the rates of operative delivery. Despite these potential limitations, the combined model proved useful in predicting adverse outcomes in the validation cohort with no significant differences between the cohort sites.

Clinical and research implications

Prediction models are aimed at helping clinicians with decision-making and are developed relatively frequently for many medical situations. Unfortunately, their usefulness largely relies on the population they are being used and it is not uncommon to observe greatly diminished predictive accuracy when a model is tested on an external cohort [30]. External validation studies are therefore crucial before the prediction models can be used in clinical practice. Unfortunately, external validation studies which adequately report the diagnostic performance measures of a model are rare [31]. Our combined model is one of the few multivariable models which aimed to predict adverse outcome in appropriate-for-gestational-age (AGA) or SGA fetuses [32-35].The combined model had an AUC of 0.72 in an external cohort without refitting which is more than the 0.70, a value considered as a fair performance indicator and is also a threshold for clinical usefulness.

Prognostic markers that are associated with adverse outcomes in SGA fetuses are useful. However, the translation of statistical findings into clinical practice is problematic due to the lack of practical tools. The IRIS mobile app, which uses our combined model algorithm, is an easy to use application that can facilitate clinical translation of our findings. Individualized risk assessment could help physicians with decision making. For example, with the use of mobile application, it would be feasible to reassure the mother of a low-risk fetus to proceed with vaginal delivery plans or to see the added risk of labor augmentation by comparing it to the baseline risk due to unmodifiable risk factors (gestational age, CPR MoM, parity).

The development of a prediction model is an arduous process and assessment of its clinical utility is an important final step [36,37]. Our prediction model can be helpful in identifying fetuses at an increased risk of intrapartum fetal compromise. Improved identification of such fetuses may reduce the incidence of fetuses born with acidemia, but may also inadvertently increase the rate of elective cesarean section. It is also possible that an increase in cesarean section rates may not

result in a decline in adverse fetal outcomes. Future studies are needed to test the utility and effectiveness of the IRIS tool mobile app.

Study strengths and limitations

We reported both the accuracy and goodness-of-fit measures as recommended by experts in study methodology [31]. Our model was aimed at a specific outcome measure as opposed to other studies which used a composite adverse outcome measure [32]. This is quite important as outcomes such as the operative delivery and neonatal unit admission which are usually blanketed under the same category, can have different prognostic and confounding variables [4,6]. The CPR values were not calculated before the analysis for this study, and therefore, limiting the effect of intervention bias on the results we have obtained. However, eliminating the intervention bias in a retrospective cohort is not possible and it is possible that our results are confounded by intervention bias. A prospective randomized trial is needed to assess the usefulness of CPR for the prevention of adverse outcomes [38]. We had a smaller number of suspected SGA fetuses in the validation cohort than the original cohort which was used to develop the prediction model. Also, we could not reach the recommended number of minimum outcomes for external validation studies (recommended 100 vs 94 current), despite including cases from two study centers. Finally, despite the results we have obtained here, this external validation study was performed on a small cohort from two institutions and we cannot confirm similar performance of the model in other populations.

Conclusion

The prediction model has modest predictive accuracy and goodness-of-fit without the need for refitting. An IRIS mobile app is available for clinicians who wish to use the predictive model in their clinical practice.

REFERENCES

1.Khalil A, Morales-Rosello J, Khan N, Nath M, Agarwal P, Bhide A, Papageorghiou A, Thilaganathan B. Is cerebroplacental ratio a marker of impaired fetal growth velocity and adverse pregnancy outcome? Am J Obstet Gynecol. 2017; 216: 606 e1- e10.

2.Khalil AA, Morales-Rosello J, Morlando M, Hannan H, Bhide A, Papageorghiou A, Thilaganathan B. Is fetal cerebroplacental ratio an independent predictor of intrapartum fetal compromise and neonatal unit admission? Am J Obstet Gynecol. 2015; 213: 54 e1-10.

3.Pilliod RA, Cheng YW, Snowden JM, Doss AE, Caughey AB. The risk of intrauterine fetal death in the small-for-gestational-age fetus. Am J Obstet Gynecol. 2012; 207: 318 e1-6.

4.Kalafat E, Morales-Rosello J, Thilaganathan B, Tahera F, Khalil A. Risk of operative delivery for intrapartum fetal compromise in small-for-gestational-age fetuses at term: an internally validated prediction model. Am J Obstet Gynecol. 2018; 218: 134 e1- e8.

5.Khalil AA, Morales-Rosello J, Elsaddig M, Khan N, Papageorghiou A, Bhide A, Thilaganathan B. The association between fetal Doppler and admission to neonatal unit at term. Am J Obstet Gynecol. 2015; 213: 57 e1-7.

6.Kalafat E, Morales-Rosello J, Thilaganathan B, Dhother J, Khalil A. Risk of neonatal care unit admission in small for gestational age fetuses at term: a prediction model and internal validation. J Matern Fetal Neonatal Med. 2018 [epub ahead of print]

7.Lindqvist PG, Molin J. Does antenatal identification of small-for-gestational age fetuses significantly improve their outcome? Ultrasound Obstet Gynecol. 2005; 25: 258-64.

8.Visentin S, Londero AP, Grumolato F, Trevisanuto D, Zanardo V, Ambrosini G, Cosmi E. Timing of delivery and neonatal outcomes for small-for-gestational-age fetuses. J Ultrasound Med. 2014; 33: 1721-8.

9.Aviram A, Yogev Y, Bardin R, Meizner I, Wiznitzer A, Hadar E. Small for gestational age newborns--does pre-recognition make a difference in pregnancy outcome? J Matern Fetal Neonatal Med. 2015; 28: 1520-4.
10.Sovio U, White IR, Dacey A, Pasupathy D, Smith GCS. Screening for fetal growth restriction with universal third trimester ultrasonography in nulliparous women in the Pregnancy Outcome Prediction (POP) study: a prospective cohort study. Lancet. 2015; 386: 2089-97.
11.Caradeux J, Eixarch E, Mazarico E, Basuki TR, Gratacos E, Figueras F. Longitudinal growth assessment for the prediction of adverse perinatal outcome in SGA-suspected fetuses. Ultrasound Obstet Gynecol. 2017. [epub ahead of print]
12.Monaghan C, Binder J, Thilaganathan B, Morales-Rosello J, Khalil A. Perinatal Loss at Term: The Role of Uteroplacental and Fetal Doppler Assessment. Ultrasound Obstet Gynecol. 2017. [epub ahead of print]
13.Gordijn SJ, Beune IM, Thilaganathan B, Papageorghiou A, Baschat AA, Baker PN, Silver RM, Wynia K, Ganzevoort W. Consensus definition of fetal growth restriction: a Delphi procedure. Ultrasound Obstet Gynecol. 2016; 48: 333-9
14.Dunn L, Sherrell H, Kumar S. Review: Systematic review of the utility of the fetal cerebroplacental ratio measured at term for the prediction of adverse perinatal outcome. Placenta. 2017; 54: 68-75.
15.Schreurs CA, de Boer MA, Heymans MW, Schoonmade LJ, Bossuyt PMM, Mol BWJ, de Groot CJM, Bax CJ. Prognostic accuracy of cerebroplacental ratio and middle cerebral artery Doppler for adverse perinatal outcomes: a systematic review and meta-analysis. Ultrasound Obstet Gynecol. 2017. [epub ahead of print]
16.Bligh LN, Alsolai AA, Greer RM, Kumar S. Pre-labour screening for intrapartum fetal compromise in low risk pregnancies at term: cerebroplacental ratio and placental growth factor. Ultrasound Obstet Gynecol. 2017. [epub ahead of print]

17. Bligh LN, Alsolai AA, Greer RM, Kumar S. Cerebroplacental ratio thresholds measured within two weeks of birth and the risk of Cesarean section for intrapartum fetal compromise and adverse neonatal outcome. Ultrasound Obstet Gynecol. 2017. [epub ahead of print]
18. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001; 54: 774-81.
19. Robinson HP, Fleming JE. A critical evaluation of sonar "crown-rump length" measurements. Br J Obstet Gynaecol. 1975; 82: 702-10.
20. Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with the use of head, body, and femur measurements--a prospective study. Am J Obstet Gynecol. 1985; 151: 333-7.
21. Bhide A, Acharya G, Bilardo CM, Brezinka C, Cafici D, Hernandez-Andrade E, Kalache K, Kingdom J, Kiserud T, Lee W, Lees C, Leung KY, Malinger G, Mari G, Prefumo F, Sepulveda W, Trudinger B. ISUOG practice guidelines: use of Doppler ultrasonography in obstetrics. Ultrasound Obstet Gynecol. 2013; 41: 233-39.
22. Morales-Rosello J, Khalil A, Morlando M, Papageorghiou A, Bhide A, Thilaganathan B. Changes in fetal Doppler indices as a marker of failure to reach growth potential at term. Ultrasound Obstet Gynecol. 2014; 43: 303-10.
23. Poon LC, Tan MY, Yerlikaya G, Syngelaki A, Nicolaides KH. Birth weight in live births and stillbirths. Ultrasound Obstet Gynecol. 2016; 48: 602-6.
24. Hadlock FP, Deter RL, Harrist RB, Park SK. Estimating fetal age: computer-assisted analysis of multiple fetal growth parameters. Radiology. 1984; 152: 497-501.
25. Monitorización fetal intraparto. Progresos de Obstetricia y Ginecología. 2005; 48: 207-16.

26.National Collaborating Centre for Women's and Children's Health commissioned by the National Institute for Health and Clinical Excellence. Intrapartum care. 2007.

27.Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med 2016; 35: 214–26.

28.R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. https:// www.R-project.org/.

29.Anim-Somuah M, Smyth RM, Jones L. Epidural versus non-epidural or no analgesia in labour. Cochrane Database Syst Rev. 2011: CD000331.

30.Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015; 68: 25-34.

31.Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu LM, Moons KG, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014; 14: 40.

32.Miranda J, Triunfo S, Rodriguez-Lopez M, Sairanen M, Kouru H, Parra-Saavedra M, Crovetto F, Figueras F, Crispi F, Gratacos E. Performance of third-trimester combined screening model for prediction of adverse perinatal outcome. Ultrasound Obstet Gynecol. 2017; 50: 353-60.

33.Triunfo S, Crispi F, Gratacos E, Figueras F. Prediction of delivery of small-for-gestational-age neonates and adverse perinatal outcome by fetoplacental Doppler at 37 weeks' gestation. Ultrasound Obstet Gynecol. 2017; 49: 364-71.

34. Valino N, Giunta G, Gallo DM, Akolekar R, Nicolaides KH. Biophysical and biochemical markers at 35-37 weeks' gestation in the prediction of adverse perinatal outcome. Ultrasound Obstet Gynecol. 2016; 47: 203-9.
35. Valino N, Giunta G, Gallo DM, Akolekar R, Nicolaides KH. Biophysical and biochemical markers at 30-34 weeks' gestation in the prediction of adverse perinatal outcome. Ultrasound Obstet Gynecol. 2016; 47: 194-202.
36. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35:1925-31.
37. van der Steeg JW, Steures P, Eijkemans MJ, Habbema JD, Bossuyt PM, Hompes PG, van der Veen F, Mol BW. Do clinical prediction models improve concordance of treatment decisions in reproductive medicine? BJOG. 2006;113:825-31.
38. Figueras F, Gratacos E, Rial M, Gull I, Krofta L, Lubusky M, Cruz-Martinez R, Cruz-Lemini M, Martinez-Rodriguez M, Socias P, Aleuanlli C, Cordero MCP. Revealed versus concealed criteria for placental insufficiency in an unselected obstetric population in late pregnancy (RATIO37): randomised controlled trial study protocol. BMJ Open. 2017; 7: e014835.

Table 1. Characteristics of the external validation cohort grouped according to the method of delivery.

	UK cohort (n=113)			**Spain cohort (n=299)**		
	Operative delivery* (n=21)	**No operative delivery* (n=92)**	**P †**	**Operative delivery* (n=73)**	**No operative delivery* (n=226)**	**P †**
Antenatal variables						
Maternal age (years)	31.0 (27.0-36.0)	29.0 (25.8-35.0)	0.212	31.0 (28.0-36.0)	32.0 (28.0-36.0)	0.973
Multiparous	4 (19.0)	40 (43.5)	0.047	14 (19.2)	112 (49.6)	<0.001
Ultrasound and Doppler variables						
Gestational age at ultrasound (weeks)	37.7 (36.7-38.4)	37.4 (36.7-38.1)	0.250	39.0 (37.7-39.6)	38.9 (38.0-39.7)	0.767
Interval between ultrasound and delivery (days)	4.0 (1.0-12.0)	7.0 (3.0-14.3)	0.055	5.0 (2.0-9.0)	8.0 (2.0-9.0)	0.642
Biparietal diameter (mm)	90.1 (88.3-93.0)	89.0 (86.5-91.0)	0.109	88.0 (86.0-90.0)	87.0 (84.0-90.0)	0.201
Biparietal diameter centile	34.1 (18.1-65.1)	28.9 (10.7-57.6)	0.411	7.3 (1.1-17.6)	3.5 (0.4-12.6)	0.130
	301.8 (286.8-313.6)	298.5 (288.4-307.5)	0.658	312.0 (302.0-320.0)	312.0 (300.0-319.0)	0.569

Abdominal circumference (mm)						
Abdominal circumference centile	0.6 (0.1-3.4)	0.7 (0.2-2.5)	0.701	1.24 (0.3-4.3)	1.1 (0.2-3.8)	0.354
Femur length (mm)	66.3 (64.3-68.0)	67.0 (65.0-69.0)	0.406	66.0 (64.0-68.0)	66.0 (64.0-68.0)	0.653
Femur length centile	0.5 (0.1-1.4)	1.7 (0.7-5.2)	0.008	0.0 (0.0-0.5)	0.1 (0.0-0.6)	0.947
Estimated fetal weight (grams)	2467.0 (2189.0-2638.0)	2377.0 (2231.0-2595.0)	0.557	2563.0 (2325.0-2714.0)	2497.0 (2327.0-2656.0)	0.340
Estimated fetal weight centile	3.7 (2.5-6.6)	6.0 (3.5-8.8)	0.092	4.2 (2.2-6.1)	3.3 (1.6-5.7)	0.090
Umbilical artery PI	0.90 (0.79-1.11)	0.89 (0.80-1.03)	0.676	0.93 (0.82-1.16)	0.88 (0.78-1.02)	0.004
Umbilical artery PI MoM	1.03 (0.90-1.28)	1.02 (0.90-1.17)	0.453	1.12 (0.96-1.36)	1.04 (0.93-1.18)	0.004
Middle cerebral artery PI	1.32 (1.07-1.44)	1.48 (1.32-1.71)	0.001	1.25 (1.09-1.50)	1.45 (1.22-1.73)	<0.001
Middle cerebral artery PI MoM	1.03 (0.96-1.17)	1.18 (1.06-1.33)	0.002	1.12 (0.95-1.33)	1.27 (1.10-1.53)	<0.001
Cerebroplacental ratio	1.43 (1.00-1.64)	1.70 (1.39-2.03)	0.004	1.36 (0.98-1.72)	1.60 (1.31-2.06)	<0.001

Cerebroplacental ratio MoM	0.77 (0.62-0.91)	0.92 (0.77-1.10)	0.008	0.77 (0.56-0.95)	0.92 (0.75-1.16)	<0.001
Intrapartum variables						
Induction of labor	18 (85.7)	71 (77.2)	0.556	52 (71.2)	120 (53.1)	0.006
Oxytocin use for labor augmentation	4 (19.0)	11 (12.0)	0.474	50 (68.5)	114 (50.4)	0.009
Epidural use	12 (57.1)	22 (23.9)	0.006	57 (78.1)	164 (72.6)	0.443
Variables at birth						
Gestational age at delivery (weeks)	38.3 (38.1-40.0)	38.9 (38.1-39.6)	0.997	39.7 (38.9-40.6)	39.9 (38.9-40.4)	0.688
Birthweight (grams)	2300.0 (2172-2600.0)	2545.0 (2295.0-2750.0)	0.060	2550.0 (2380.0-2750.0)	2705.0 (2450.0-2870.0)	0.003
Birthweight centile	1.5 (0.6-3.2)	3.4 (1.5-7.6)	0.014	1.9 (0.8-3.4)	4.1 (1.5-8.3)	<0.001
Small for gestational age	21 (100.0)	78 (84.8)	0.068	68 (93.2)	179 (79.2)	0.006
Neonatal care unit admission	2 (9.5)	5 (5.4)	0.617	12 (16.4)	19 (8.4)	0.035

*For presume fetal compromise

†Group comparisons were made with either t-test, Mann-Whitney-U or Fisher's exact test.

MoM: multiple of median, PI: pulsatility index

Data provided as median and interquartile range (IQR) or number (percentage).

Table 2. The parameter estimates of prediction models for the original and external validation cohorts.

	Odds ratio, 95% CI (Original prediction model)*	**p-value**	**Odds ratio, 95% CI (Refitted model for validation cohort)***	**p-value**
Antenatal model variables				
Intercept	0.69 (0.38-1.26)	0.239	2.58 (1.10-6.21)	0.042
Multiparity	0.27 (0.17-0.41)	<0.001	0.27 (0.15-0.48)	<0.001
Abdominal circumference centile	0.96 (0.93-0.99)	0.027	1.02 (0.98-1.07)	0.320
Gestation over 39 weeks' at delivery	1.97 (1.36-2.90)	<0.001	1.40 (0.83-2.40)	0.157
Cerebroplacental ratio MoM	0.33 (0.16-0.66)	0.002	0.09 (0.03-0.25)	<0.001
Combined model variables				
Intercept	0.19 (0.09-0.40)	<0.001	2.74 (0.45-16.26)	0.266
Augmentation of labor	3.09 (1.60-5.90)	<0.001	3.16 (1.19-8.79)	0.022
Induction of labor	2.26 (1.44-3.59)	<0.001	3.06 (1.32-7.68)	0.012
Epidural analgesia	2.73 (1.89-3.94)	<0.001	0.31 (0.05-1.87)	0.196
Gestation over 39 weeks' at delivery	1.65 (1.12-2.46)	0.011	1.42 (0.82-2.51)	0.217
Cerebroplacental ratio MoM	0.35 (0.16-0.72)	0.005	0.10 (0.04-0.28)	<0.001
Multiparity	0.36 (0.23-0.56)	<0.001	0.32 (0.17-0.57)	<0.001

Interaction term for induction of labor and augmentation using oxytocin	0.43 (0.19-0.95)	0.037	0.30 (0.09-0.94)	0.041
Cohort location				
-Spain	NA	NA	Reference	
-UK	NA	NA	0.40 (0.14-1.08)	0.075
Interaction term for epidural and cohort location	NA	NA	3.07 (0.88-10.9)	0.079

*Parameter estimates were obtained via generalized linear models using a logit link.

CI: confidence interval, MoM: multiple of median

Table 3. Diagnostic accuracy parameters of the original and refitted combined model for different risk cut-offs.

Diagnostic accuracy parameters	Original combined model	Refitted combined model	Posterior probability†
Risk cut-off >10%			
- Accuracy*	44.4% (39.6-49.4%)	50.2% (45.3-55.2%)	95.1%
-Sensitivity	95.7%	95.7%	50.3%
-Specificity	29.3%	36.8%	80.8%
Risk cut-off >20%			
- Accuracy	57.3% (52.4-62.1%)	63.3% (58.5-68.0%)	96.7%
-Sensitivity	81.9%	72.3%	5.6%
-Specificity	50.0%	60.7%	99.7%
Risk cut-off >30%			
- Accuracy	67.2% (62.5-71.8%)	73.3% (68.8-77.5%)	97.1%
-Sensitivity	63.8%	56.4%	14.2%
-Specificity	68.2%	78.3%	99.8%
Risk cut-off >40%			
- Accuracy	70.9% (66.2-75.2%)	76.9% (72.6-80.9%)	97.6%
-Sensitivity	43.6%	40.4%	34.7%
-Specificity	78.9%	87.7%	96.2%
Risk cut-off >50%			
- Accuracy	76.0% (71.2-80.0%)	78.6% (74.4-82.5%)	82.3%
-Sensitivity	22.3%	22.3%	50.8%
-Specificity	91.8%	95.3%	96.3%

Percentages are given as mean estimates and 95% confidence intervals within the brackets

* The confidence intervals for the accuracy estimates were calculated with 10.000 stratified bootstrap replicates

†Posterior probabilities of improvement in diagnostic accuracy via model refitting were calculated using a Bayesian framework(Likehood~Binomial(n,p); Prior~Beta(0.5,0.5))

CI: confidence interval, FPR: false positive rate

Figure 1. Patient enrollment flow-chart.

Figure 2. Receiver operating characteristics curves for the antenatal model using the original (dashed lines) and re-estimated coefficients (straight line). The De Long test indicated statistically non-significant improvement in the model accuracy with refitting (AUC 0.73, 95% CI: 0.67-0.79 vs AUC 0.70, 95% CI: 0.64-0.76, refitted and original antenatal model respectively, P=0.076)

Figure 3. Receiver operating characteristics curves for the combined model using the original (dashed lines) and re-estimated coefficients (straight line). The De Long test indicated a statistically non-significant improvement in the model accuracy with refitting (AUC 0.76, 95% CI: 0.70-0.81 vs AUC 0.72, 95% CI: 0.67-0.77, refitted and original combined model respectively, P=0.060)

Supplement Figure 1. The calibration plot for the antenatal model using the original regression coefficients. The black line represents the predicted means and yellow area represents the confidence intervals. Deviation from the redline indicates predicted and observed averages are incongruent. The Hosmer-Lemeshow test indicated no significant effect for poor-fit (P=0.464).

Supplement Figure 2. The calibration plot for the antenatal model using the re-estimated regression coefficients. The black line represents the predicted means and yellow area represents the confidence intervals. Deviation from the redline indicates predicted and observed averages are incongruent. The Hosmer-Lemeshow test indicated a significant effect for poor-fit (P=0.008).

Supplement Figure 3. The calibration plot for the combined model using the original regression coefficients. The black line represents the predicted means and yellow area represents the confidence intervals. Deviation from the redline indicates predicted and observed averages are incongruent. The Hosmer-Lemeshow test indicated no significant effect for poor-fit (P=0.268).

Supplement Figure 4. The calibration plot for the combined model using the re-estimated regression coefficients. The black line represents the predicted means and yellow area represents the confidence intervals. Deviation from the

redline indicates predicted and observed averages are incongruent. The Hosmer-Lemeshow test indicated no significant effect for poor-fit (P=0.545).

Supplement Figure 5. Receiver operating characteristics curves for the antenatal (a) and combined (b) model for the Spanish cohort (straight lines) and UK cohort (dashed lines) using the original model coefficients. There were no significant differences between cohort locations regarding predictive accuracy for the combined model (Area under the curve: 0.71 vs 0.73, Spain and England, respectively. P=0.755) and the antenatal model (Area under the curve: 0.70 vs 0.66, Spain and UK, respectively. P=0.531).