Ultrasound Obstet Gynecol 2022; **59**: 209–219
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.23757.

External validation of prognostic models to predict stillbirth using International Prediction of Pregnancy Complications (IPPIC) Network database: individual participant data meta-analysis

J. ALLOTEY[1,2#], R. WHITTLE[3#], K. I. E. SNELL[3], M. SMUK[4], R. TOWNSEND[5,6], P. VON DADELSZEN[7], A. E. P. HEAZELL[8], L. MAGEE[7], G. C. S. SMITH[9], J. SANDALL[7,10], B. THILAGANATHAN[5,6], J. ZAMORA[1,11,12], R. D. RILEY[3], A. KHALIL[5,6] and S. THANGARATINAM[1,13], on behalf of the IPPIC Collaborative Network*

[1]*WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK;* [2]*Institute of Applied Health Research, University of Birmingham, Birmingham, UK;* [3]*Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK;* [4]*Medical Statistics Department, London School of Hygiene and Tropical Medicine, London, UK;* [5]*Fetal Medicine Unit, St George's University Hospitals NHS Foundation Trust, University of London, London, UK;* [6]*Vascular Biology Research Centre, Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK;* [7]*Department of Women and Children's Health, School of Life Course Sciences, King's College London, London, UK;* [8]*Maternal and Fetal Health Research Centre, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK;* [9]*Department of Obstetrics and Gynaecology, NIHR Biomedical Research Centre, Cambridge University, Cambridge, UK;* [10]*Health Service and Population Research Department, Centre for Implementation Science, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK;* [11]*Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal (IRYCIS), Madrid, Spain;* [12]*CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain;* [13]*Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK*

KEYWORDS: external validation; individual participant data; intrauterine death; prediction model; stillbirth

CONTRIBUTION

What are the novel findings of this work?
We identified 40 published stillbirth prediction models. The full model equation was reported for only eight models, of which external validation of model performance using individual patient data from the International Prediction of Pregnancy Complications (IPPIC) Network database was possible for three. All three models generally had poor summary discrimination and calibration, with little to no clinical value for decision-making.

What are the clinical implications of this work?
None of the externally validated stillbirth prediction models can be recommended for use in clinical practice. Further research is needed to further validate these and other models, identify stronger prognostic factors and develop more robust prediction models.

ABSTRACT

Objective *Stillbirth is a potentially preventable complication of pregnancy. Identifying women at high risk of stillbirth can guide decisions on the need for closer surveillance and timing of delivery in order to prevent fetal death. Prognostic models have been developed to predict the risk of stillbirth, but none has yet been validated externally. In this study, we externally validated published prediction models for stillbirth using individual participant data (IPD) meta-analysis to assess their predictive performance.*

Methods *MEDLINE, EMBASE, DH-DATA and AMED databases were searched from inception to December 2020 to identify studies reporting stillbirth prediction models. Studies that developed or updated prediction models for stillbirth for use at any time during pregnancy were included. IPD from cohorts within the International*

[Correction added on 29 April 2022, after first online publication: In supporting information, Appendix S1 was corrected.]

Correspondence to: Dr J. Allotey, Room 14, 4th Floor East, Institute of Translational Medicine, Heritage Building, Mindelsohn Way, Edgbaston, Birmingham B15 2TH, UK (e-mail: j.allotey.1@bham.ac.uk)

#J.A. and R.W. are joint first authors.

*Members of the IPPIC Collaborative Network are listed in Appendix S1.

Accepted: 2 August 2021

Prediction of Pregnancy Complications (IPPIC) Network were used to validate externally the identified prediction models whose individual variables were available in the IPD. The risk of bias of the models and cohorts was assessed using the Prediction study Risk Of Bias ASsessment Tool (PROBAST). The discriminative performance of the models was evaluated using the C-statistic, and calibration was assessed using calibration plots, calibration slope and calibration-in-the-large. Performance measures were estimated separately in each cohort, as well as summarized across cohorts using random-effects meta-analysis. Clinical utility was assessed using net benefit.

Results *Seventeen studies reporting the development of 40 prognostic models for stillbirth were identified. None of the models had been previously validated externally, and the full model equation was reported for only one-fifth (20%, 8/40) of the models. External validation was possible for three of these models, using IPD from 19 cohorts (491 201 pregnant women) within the IPPIC Network database. Based on evaluation of the model development studies, all three models had an overall high risk of bias, according to PROBAST. In the IPD meta-analysis, the models had summary C-statistics ranging from 0.53 to 0.65 and summary calibration slopes ranging from 0.40 to 0.88, with risk predictions that were generally too extreme compared with the observed risks. The models had little to no clinical utility, as assessed by net benefit. However, there remained uncertainty in the performance of some models due to small available sample sizes.*

Conclusions *The three validated stillbirth prediction models showed generally poor and uncertain predictive performance in new data, with limited evidence to support their clinical application. The findings suggest methodological shortcomings in their development, including overfitting. Further research is needed to further validate these and other models, identify stronger prognostic factors and develop more robust prediction models. © 2021 The Authors. Ultrasound in Obstetrics & Gynecology published by John Wiley & Sons Ltd on behalf of International Society of Ultrasound in Obstetrics and Gynecology.*

INTRODUCTION

Stillbirth continues to be a major burden globally, accounting for almost two-thirds of perinatal mortality[1,2]. In the UK, the stillbirth rate was largely unchanged from 2000 to 2015, and, in 2017, the rate was one of the highest in Europe, at 4.2 stillbirths/1000 births[3–5]. Prediction and individualization of risk remain key priorities for stillbirth research[6,7], because accurate identification of women at high risk of stillbirth can guide decisions on the need for closer surveillance and timing of delivery in order to prevent fetal death. A recent review that identified existing prediction models for stillbirth reported that none had been validated externally[8]. As a result, no stillbirth prediction model is used routinely in clinical practice, and none has been recommended by any national or international guidelines.

Independent, external validation and comparison of existing multivariable stillbirth prediction models is important to help identify which prediction model (if any) performs best and is potentially applicable in clinical practice. However, the relative rarity of this devastating outcome limits rigorous investigation of existing stillbirth prediction models in single-cohort studies. An individual participant data (IPD) meta-analysis combining the raw data from multiple studies has great potential for use in validating externally existing models by increasing the sample size beyond what is feasible in a single study, thereby increasing the number of events observed[9–12]. It would also allow evaluation of the generalizability and transportability of the predictive performance of the models across a range of clinical settings being considered for their application.

We therefore set out to identify, appraise critically and validate externally existing multivariable prognostic models for stillbirth prediction using IPD meta-analysis within the independent International Prediction of Pregnancy Complications (IPPIC) Network database, and to assess the clinical utility of the models using decision-curve analysis (DCA).

METHODS

This study was based on a prospective protocol registered in the international prospective register of systematic reviews (PROSPERO; registration number: CRD42018074788), and is reported in line with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) recommendations for reporting risk prediction model validation studies[13]. Ethics approval was not required, as the study involved secondary analysis of existing anonymized data.

Literature search and selection of prediction models for external validation using IPPIC Network database

MEDLINE, EMBASE, DH-DATA and AMED databases were searched systematically, from inception to December 2020, to identify all studies that developed or updated prognostic models for stillbirth for use at any time during pregnancy. We also searched manually reference lists of relevant articles and systematic reviews to identify potentially eligible studies. The search included terms for stillbirth, intrauterine fetal death and perinatal mortality, and study selection was performed independently by two researchers (J.A. and R.T.). The complete search strategy is provided in Appendix S2.

Stillbirth model eligibility criteria, data extraction and risk-of-bias assessment

Studies that reported the development or update of a multivariable model with at least three variables to

predict the risk of stillbirth in pregnant women and that reported the model equation in the publication were included. No attempts were made to contact authors of studies that did not report the model equation. Given the wide international variation in the definition of stillbirth, we accepted the authors' definition of stillbirth (ante- or intrapartum fetal death) and included models developed for use at any time in pregnancy. We excluded models that predicted stillbirth as part of a composite adverse outcome, those that contained predictors not measured in any of the cohorts within the IPPIC Network database and those for which too few outcome events for external validation (< 10 stillbirths) were reported in IPPIC Network cohorts that contained the same predictors as the model.

We extracted data on the definition of stillbirth, number of participants and events, population type, predictors in the final model and reported model performance. Based on information in the model development studies, we assessed the risk of bias of included models using the Prediction study Risk Of Bias ASsessment Tool (PROBAST)[14], across the four domains of participant selection, predictors, outcome and analysis. Risk of bias was assessed independently by two researchers (J.A. and R.T.). Disagreements were resolved by discussion with a third researcher (S.T.). Risk of bias was classified as low, high or unclear for each domain, as well as overall. Each domain included signaling questions rated as yes, probably yes, probably no, no or no information. Domains with any signaling question rated as probably no or no were considered to have potential for bias and classed as high risk. The overall risk of bias was considered to be low if the risk of bias was classified as low in all domains, high if any one domain had a high risk of bias and unclear for any other classifications.

IPPIC Network

Cohorts for inclusion in the IPPIC Network database were identified by reviewing systematically the literature on the risk of pregnancy complications, including pre-eclampsia, stillbirth and fetal growth restriction (FGR), and the research groups who had undertaken the primary studies were invited to join the IPPIC Network and share their primary IPD. In addition, major databases and repositories were searched and researchers within the IPPIC Network were contacted to identify relevant studies or datasets that may have been missed, including unpublished research and birth cohorts. The datasets were formatted, cleaned and harmonized, and the quality of each cohort was assessed using the participants, predictors and outcome domains of the PROBAST tool[14]. The study population could vary from low to high risk of development of complications. The IPPIC Network includes nearly 150 collaborators from 26 countries, contributing IPD for over 4 million pregnancies, and contains data on maternal characteristics, obstetric history, clinical assessment and tests, as well as various maternal and offspring outcomes. The database is a living repository and is enriched regularly with additional studies. We consider the predictor variables contained within the IPPIC Network database to represent measures which are easy to obtain in a clinical setting, reflecting their availability in routine practice. Methods on how cohorts within the IPPIC Network database were identified and harmonized have been published previously[15–17].

Statistical analysis for external validation using IPPIC Network database

Data harmonization and set-up

Predictors or outcomes of existing prediction models that were missing partially for < 95% of individuals in any cohort were imputed multiply under the missing-at-random assumption, using multiple imputation by chained equations[18,19]. Linear regression was used for imputation of approximately normally distributed continuous variables, logistic regression for binary variables and multinomial logistic regression for categorical variables with more than two categories. Multiple imputation was carried out for each individual cohort separately and generated 50 imputed datasets for each. Other predictors that were available within the cohort as auxiliary variables were also included in the imputation models. Imputation checks were completed by evaluating histograms, summary statistics and tables of values across imputations, as well as checking trace plots for convergence issues.

External validation of models

Each model was validated by applying the model equation to each participant in the cohort to calculate the linear predictor for that participant (LP_i, value of the linear combination of predictors in the model equation for individual i), as well as the predicted probability of stillbirth (inverse logit transformation of LP_i). For each prediction model, the distribution of LP_i values was summarized for each cohort, and performance statistics were calculated in each imputed dataset and then averaged across imputations using Rubin's rules to obtain one estimate and standard error for each performance statistic in each cohort[20].

The discriminative performance of the models was assessed using the *C*-statistic (summarized as the area under the receiver-operating-characteristics curve, where 1 indicates perfect discrimination and 0.5 indicates no discrimination beyond chance), and calibration was assessed using calibration slope (slope of the regression line fitted between predicted and observed risk probabilities on the logit scale, with 1 being the ideal value) and calibration-in-the-large (the extent to which model predictions are systematically too low or too high across the cohort, ideal value of 0)[21,22]. Model calibration was also assessed visually in cohorts with at least 100 events, using calibration plots representing the average predicted probability for risk

groups categorized using deciles of predicted probability against the observed proportion in each group. A locally weighted scatterplot smoother (LOWESS) curve was applied to show calibration across the entire range of predicted probabilities at the individual level (i.e. without categorization). For the calibration plots, average predicted probabilities were obtained for individuals by pooling their linear predictor values across imputed datasets using Rubin's rules and then transforming to the probability scale.

Performance measures of prediction models that were validated in more than two independent cohorts were summarized using random-effects meta-analysis to calculate a summary estimate for the model's discriminative performance and calibration. Each model performance statistic was summarized as the average and 95% CI, calculated using the Hartung–Knapp–Sidik–Jonkman approach[23,24]. Between-study heterogeneity (τ^2) and the proportion of variability due to between-study heterogeneity (I^2)[25] were summarized.

Decision-curve analysis

We performed DCA to assess the clinical value of the models in cohorts with at least 100 events. This analysis allowed us to determine the net benefit of the models across a range of clinically plausible threshold probabilities (which included any values up to 0.1, given the general very low risk of stillbirth), compared with either simply classifying all women as having the outcome or no women as having the outcome[26]. The strategy with the highest net benefit at a particular threshold has the highest clinical value[27]. The net benefit is represented as a function of the decision threshold on decision-curve plots.

All statistical analyses were performed using Stata, version 15 (StataCorp. LLC, College Station, TX, USA).

RESULTS

From 5055 citations, 17 articles describing the development of 40 stillbirth prediction models, published between 2007 and 2020, were identified (Table S1). The full model equation was reported for only eight (20%) models. There were three studies (Smith *et al.*[28] (Smith), Yerlikaya *et al.*[29] (Yerlikaya) and Trudell *et al.*[30] (Trudell)) reporting three prediction models meeting our inclusion criteria for external validation in IPD from the IPPIC Network database (Figure 1).

Characteristics of included models

The characteristics of the included studies and models are described in Table 1. All three models were developed using binary logistic regression in unselected populations of pregnant women[28–30], and the definition of stillbirth varied between the studies. Two models included only maternal clinical characteristics as predictors[29,30], while one model additionally included an ultrasound marker[28]. Only one study had at least 10 events per predictor for model development[29], while the others did not justify whether their sample size was sufficient. Using the PROBAST tool, the overall risk of bias for all three models was high, with all models assessed as being at high risk of bias in the analysis domain.

Characteristics of IPPIC validation cohorts

Of the 78 cohorts in the IPPIC data repository, 19 (24%; 491 201 pregnant women) contained relevant data that could be used to validate externally at least one of three prediction models identified. Only women with singleton pregnancy were used for external validation. Maternal characteristics and outcomes of pregnancies in the IPPIC Network cohorts are summarized in Table 2. The prevalence of stillbirth $\geq$ 24 weeks' gestation ranged from 0.1% to 1.2%. One-quarter of the studies (26%, 5/19) included only low-risk women, while one-fifth (21%, 4/19) included only high-risk women. Seventy-four percent (14/19) of the cohorts had an overall low risk of bias, 21% (4/19) had a high risk and one had an unclear risk, as assessed by PROBAST (Table S2). The proportion of cases with missing data for each predictor and outcome is shown in Table S3.

External validation and meta-analysis of predictive performance

The Smith model was validated in three cohorts, the Yerlikaya model in four cohorts and the Trudell model in 17 cohorts. Two of the cohorts used to validate the Smith model and all four of the cohorts used to validate the Yerlikaya model were also used to validate the Trudell model. Direct comparison of the performance of the prediction models was not possible due to differences in the outcome of each model. The distributions of the linear predictor and predicted probability for each model and validation cohort are shown in Table S4.

Model predictive performance

In the different validation cohorts, the *C*-statistic ranged from 0.56 to 0.82 for the Smith model, from 0.54 to 0.73 for the Yerlikaya model and from 0.34 to 0.69 for the Trudell model (Table 3). The Trudell model had the lowest overall discrimination across the validation cohorts. The summary *C*-statistic of the model was 0.65 (95% CI, 0.53–0.75) for the Smith model, 0.61 (95% CI, 0.43–0.77) for the Yerlikaya model and 0.53 (95% CI, 0.51–0.55) for the Trudell model (Table 3). The 95% CIs for the Smith and Yerlikaya models were wide due to the lower number of cohorts available for their validation.

Calibration statistics for each model in the different validation cohorts are shown in Table 3. Summary calibration slopes were < 1 for all models, indicative

Table 1 Characteristics of stillbirth prediction models included in external validation study

Variable	*Smith*[28]	*Yerlikaya*[29]	*Trudell*[30]
Year	2007	2016	2017
Country	UK	UK	USA
Population	Pregnant women at 22–24 w, excluding those with short cervix, from seven hospitals	Women with singleton pregnancy at 11–25 w, attending two hospitals for routine pregnancy care	Women with singleton pregnancy in second trimester, attending for routine anatomical screening
Women (*n*)	30 519	113 415	57 326
Candidate predictors (*n*)	17	17	NR
Predictors included in model	UtA-PI, BMI (kg/m^2), ethnicity	Weight (kg), ethnicity, assisted conception, smoking, hypertension, APS, SLE, diabetes, previous stillbirth	MA (years), ethnicity, parity, BMI (kg/m^2), smoking, hypertension, diabetes
Prediction model equation for LP*	LP = −7.806 + 0.867 (mean UtA-PI) + 0.768 (if BMI 25–29.9) + 0.768 (if BMI ≥ 30) + 0.624 (if African-American)	LP = −6.02615 + 0.01037 (weight − 69) + 0.70027 (if Afro-Caribbean) + 0.57994 (if assisted conception) + 0.53367 (if smokes cigarettes) + 0.96253 (if chronic hypertension) + 1.28416 (if APS or SLE) + 0.93628 (if diabetic) + 1.57086 (if parous with previous stillbirth)	LP = −6.8772 − 0.8707 (if MA < 18) + 0.2094 (if MA 35–39) + 0.4377 (if MA > 40) + 0.8536 (if Black) + 0.3423 (if nulliparous) − 0.0219 (if BMI 25–29.9) + 0.5607 (if BMI 30–34.9) − 0.5948 (if BMI 35–39.9) + 0.1593 (if BMI > 40) + 0.2770 (if current smoker) + 0.6255 (if chronic hypertension) + 0.9863 (if pregestational diabetes)
Outcome	Stillbirth ≥ 33 w	Stillbirth ≥ 24 w	Stillbirth ≥ 32 w
Events (*n*)	109	396	330
Discrimination AUC (95% CI)	0.67 (0.60–0.75)	0.64 (0.61–0.67)	0.66 (0.60–0.72)
PROBAST RoB	High	High	High

Only first author of each study is given. *For logistic regression, logit (p) = LP, where linear predictor (LP) = $\alpha + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots$, and absolute predicted probabilities (p) can be obtained using transformation $p = \frac{e^{LP}}{1+e^{LP}}$. APS, antiphospholipid syndrome; AUC, area under the receiver-operating-characteristics curve; BMI, body mass index; MA, maternal age; NR, not reported; PROBAST, Prediction study Risk Of Bias ASsessment Tool; RoB, risk of bias; SLE, systemic lupus erythematosus; UtA-PI, uterine artery pulsatility index; w, weeks' gestation.

of overfitting during model development. In particular, the 95% CIs for the calibration slope were below 1 for both the Yerlikaya and Trudell models, indicating extreme risk predictions compared with the observed risk (Table 3).

Each of the three models was validated in a cohort with at least 100 events. The calibration plots showed miscalibration of the predicted risk of stillbirth for all three models (Figure 2). However, the predicted probabilities were all less than 0.02; therefore, absolute-risk differences remain small. The 95% CI was wide for the calibration slope of the Smith model, due to a lower number of events in the validation cohorts available for this model. Further investigation is therefore required for this model.

Net benefit of model use

DCA for all three models in cohorts with at least 100 events showed little or no improvement in the net benefit at any probability threshold compared to a treat-all or treat-none strategy (Figure 3).

DISCUSSION

Summary of findings

Only one-fifth of published stillbirth prognostic models reported the model equation required for independent external validation. External validation using IPD from cohorts included in the IPPIC Network data repository was possible for three models, all of which were developed in high-income countries. The models were developed mostly using maternal clinical characteristics, but one model additionally included an ultrasound marker. Assessment of the risk of bias of the original model development studies using PROBAST suggested risk of bias concerns, and IPD meta-analysis of model performance showed low discriminative ability and poor calibration, with summary calibration slopes < 1, indicative of overfitting during model development. The models had no clinical utility, as assessed by DCA. Although each of the three models could be validated in a cohort with at least 100 events, CIs of predictive performance were wide for the Smith model,

Table 2 Maternal characteristics and outcomes in International Prediction of Pregnancy Complications (IPPIC) Network cohorts used for external validation of stillbirth prediction models

							Outcome		
Cohort	*Women*	*Population type*	*MA (years) (mean ± SD (range))*	*BMI (kg/m²) (median [IQR] (range))*	*White ethnicity*	*Nulliparous*	*Stillbirth ≥ 24 weeks*	*Stillbirth ≥ 32 weeks*	*Stillbirth ≥ 33 weeks*
St George's[42]	54 635	Mixed	30.5 ± 5.6 (13–54)	23.5 [21.3–26.8] (13–54)	33 257 (61)	29 313 (54)	233 (0.43)	160 (0.29)	148 (0.27)
TEST[43]	557	Low risk	32.0 ± 4.8 (18–43)	24.0 [21.6–27.1] (17.4–45.2)	539 (97)	557 (100)	5 (0.90)	4 (0.72)	4 (0.72)
POP[44]	4212	Mixed	29.9 ± 5.1 (16–48)	24.1 [21.8–27.3] (14.7–54.7)	3900 (93)	4212 (100)	11 (0.26)	8 (0.19)	8 (0.19)
Allen[45]	1045	Mixed	29.9 ± 5.1 (15–48)	23.6 [21.0–26.8] (14.8–51.1)	398 (38)	584 (56)	3 (0.29)	3 (0.29)	3 (0.29)
Goetzinger[46]	4035	Mixed	34.8 ± 4.4 (16–52)	24.4 [21.8–28.8] (15.4–62.4)	3282 (81)	751 (19)	15 (0.37)	15 (0.37)	15 (0.37)
JSOG[47]	379 390	Mixed	32.2 ± 5.4 (10–59)	20.5 [19.0–22.6] (10.5–69.8)	0 (0)	195 983 (52)	1792 (0.47)	895 (0.24)	801 (0.21)
StorkG[48]	812	Mixed	29.8 ± 4.8 (19–45)	25.1 [22.3–28.4] (16.2–49.8)	375 (46)	377 (46)	6 (0.74)	5 (0.62)	4 (0.49)
SCOPE[49]	5628	Low risk	28.7 ± 5.5 (14–45)	24.2 [21.9–27.5] (15.4–58.5)	5061 (90)	5628 (100)	17 (0.30)	9 (0.16)	8 (0.14)
ALSPAC[50]	15 038	Mixed	27.7 ± 4.9 (13–46)	21.5 [19.7–23.7] (11.7–61.3)	11 769 (78)	5704 (38)	41 (0.27)	27 (0.18)	26 (0.17)
Antsaklis[51]	3328	Low risk	30.9 ± 4.8 (14–47)	22.7 [20.6–25.7] (14.5–50.1)	3229 (97)	3328 (100)	2 (0.06)	2 (0.06)	2 (0.06)
WHO[52]	7273	High risk	22.5 ± 5.8 (11–51)	23.1 [21.0–26.1] (13.5–54.8)	2222 (31)	6710 (92)	8 (0.11)	8 (0.11)	8 (0.11)
Andersen[53]	2120	Mixed	30.2 ± 4.5 (17–45)	23.4 [21.2–26.2] (14.9–49.9)	1765 (83)	1193 (56)	6 (0.28)	4 (0.19)	4 (0.19)
NICHD HR[54]	1848	High risk	27.1 ± 6.3 (15–43)	28.4 [23.5–35.0] (13.4–68.5)	612 (33)	430 (23)	23 (1.24)	8 (0.43)	8 (0.43)
NICHD LR[55]	3097	Low risk	20.6 ± 4.4 (15–39)	22.7 [20.4–25.7] (13.4–51.2)	548 (18)	3097 (100)	13 (0.42)	6 (0.19)	6 (0.19)
POUCH[56]	3019	Mixed	26.4 ± 5.8 (15–47)	27.7 [24.3–32.9] (15.1–66.3)	2018 (67)	1293 (43)	10 (0.33)	4 (0.13)	4 (0.13)
Rumbold[57]	1877	Low risk	26.4 ± 5.7 (13–44)	24.1 [21.5–27.6] (13.7–57.6)	1777 (95)	1877 (100)	11 (0.59)	9 (0.48)	9 (0.48)
Indonesian cohort[58]	2223	Mixed	28.6 ± 5.9 (10–59)	22.9 [20.1–26.3] (13.3–67.6)	0 (0)	664 (30)	12 (0.54)	6 (0.27)	6 (0.27)
van Oostwaard 2012[59]	425	High risk	32.0 ± 4.1 (23–42)	24.3 [21.5–27.9] (16.2–41.8)	288 (68)	0 (0)	2 (0.47)	2 (0.47)	2 (0.47)
Van Oostwaard 2014[60]	639	High risk	32.1 ± 4.4 (21–43)	25.9 [22.5–31.2] (17.7–56.5)	360 (56)	0 (0)	5 (0.78)	3 (0.47)	3 (0.47)

Data are given as *n* or *n* (%), unless stated otherwise. BMI, body mass index; IQR, interquartile range; MA, maternal age.

Table 3 Individual and summary performance statistics of stillbirth prediction models in International Prediction of Pregnancy Complications (IPPIC) Network cohorts used for external validation

			Performance statistic (95% CI) and heterogeneity [I^2; τ^2]		
Model	*Women*	*Events*	*C-statistic*	*Calibration slope*	*Calibration-in-the-large*
Smith (2007)[28] *					
St George's[42]	54 635	148 (0.27)	0.65 (0.60–0.70)	0.87 (0.57 to 1.16)	0.57 (0.41 to 0.73)
TEST[43]	557	4 (0.72)	0.82 (0.52–0.95)	1.57 (0.16 to 2.99)	1.74 (0.75 to 2.72)
POP[44]	4212	8 (0.19)	0.56 (0.36–0.75)	0.49 (−0.93 to 1.92)	0.29 (−0.41 to 0.98)
Summary	59 404	160 (0.27)	0.65 (0.53–0.75) [0%; 0]	0.88 (0.26 to 1.50) [0%; 0]	0.76 (−0.95 to 2.48) [76.6%; 0.292]
Yerlikaya (2016)[29] †					
Allen[45]	1045	3 (0.29)	0.64 (0.31–0.88)	0.54 (−1.57 to 2.65)	−1.52 (−2.66 to −0.39)
Goetzinger[46]	4035	15 (0.37)	0.63 (0.42–0.80)	0.66 (−0.10 to 1.42)	−1.98 (−2.37 to −1.59)
JSOG[47]	379 390	1792 (0.47)	0.54 (0.53–0.56)	0.44 (0.32 to 0.55)	−0.74 (−0.79 to −0.70)
StorkG[48]	812	6 (0.74)	0.73 (0.56–0.85)	1.04 (−0.42 to 2.50)	−0.41 (−1.15 to 0.34)
Summary	385 282	1816 (0.47)	0.61 (0.43–0.77) [48.6%; 0.102]	0.45 (0.26 to 0.63) [0%; 0]	−1.15 (−2.35 to 0.05) [91.4%; 0.462]
Trudell (2017)[30] ‡					
SCOPE[49]	5628	9 (0.16)	0.34 (0.20–0.51)	−1.84 (−3.77 to 0.86)	−0.03 (−0.69 to 0.62)
Allen[45]	1045	3 (0.29)	0.47 (0.18–0.79)	−0.28 (−3.43 to 2.87)	0.58 (−0.56 to 1.71)
ALSPAC[50]	15 038	27 (0.18)	0.48 (0.33–0.63)	−0.04 (−1.77 to 1.68)	0.15 (−0.23 to 0.53)
Goetzinger[46]	4035	15 (0.37)	0.54 (0.27–0.79)	0.52 (−0.70 to 1.75)	1.20 (0.78 to 1.62)
Antsaklis[51]	3328	2 (0.06)	0.43 (0.10–0.84)	−1.08 (−4.72 to 2.57)	−1.27 (−2.64 to 0.10)
WHO[52]	7273	8 (0.11)	0.54 (0.40–0.67)	0.17 (−0.73 to 1.07)	1.73 (1.00 to 2.46)
Andersen[53]	2120	4 (0.19)	0.62 (0.28–0.87)	1.55 (−2.00 to 5.10)	0.25 (−0.73 to 1.23)
NICHD HR[54]	1848	8 (0.43)	0.61 (0.39–0.80)	0.44 (−0.57 to 1.44)	−0.03 (−0.72 to 0.67)
NICHD LR[55]	3097	6 (0.19)	0.64 (0.35–0.85)	0.88 (−0.60 to 2.36)	0.05 (−0.76 to 0.85)
POUCH[56]	3019	4 (0.13)	0.64 (0.42–0.81)	0.66 (−1.10 to 2.42)	−0.38 (−1.36 to 0.60)
Rumbold[57]	1877	9 (0.48)	0.47 (0.27–0.69)	−0.68 (−2.64 to 1.28)	1.07 (0.42 to 1.73)
JSOG[47]	379 390	895 (0.24)	0.53 (0.51–0.55)	0.41 (0.18 to 0.65)	0.49 (0.43 to 0.56)
Indonesian cohort[58]	2223	6 (0.27)	0.69 (0.48–0.85)	1.92 (0.07 to 3.78)	1.30 (0.57 to 2.02)
StorkG[48]	812	5 (0.62)	0.43 (0.16–0.76)	0.29 (−1.79 to 2.37)	1.58 (0.77 to 2.39)
van Oostwaard 2012[59]	425	2 (0.47)	0.64 (0.35–0.86)	0.65 (−0.75 to 2.05)	2.89 (1.71 to 4.06)
Van Oostwaard 2014[60]	639	3 (0.47)	0.59 (0.24–0.87)	0.38 (−1.48 to 2.24)	1.20 (0.03 to 2.37)
POP[44]	4212	8 (0.19)	0.63 (0.40–0.82)	1.20 (−0.42 to 2.81)	0.09 (−0.61 to 0.78)
Summary	436 009	1014 (0.23)	0.53 (0.51–0.55) [0%; 0]	0.40 (0.19 to 0.62) [0%; 0]	0.64 (0.18 to 1.11) [89.1%; 0.552]

Only first author of each study is given. Data are given as *n* or *n* (%), unless stated otherwise. Outcomes were: stillbirth *≥ 33, †≥ 24 and ‡≥ 32 weeks' gestation.

suggesting that further validation is needed for this model.

Strengths and limitations

To our knowledge, this is the first systematic review and external validation study of stillbirth prediction models[8,31]. This study, with its large sample size, allowed for the evaluation of the predictive performance of each model across multiple cohorts, as well as the overall performance, using IPD meta-analysis. We used multiple imputation of predictors and outcomes for each cohort separately to avoid loss of useful information and ensure that we did not mask any heterogeneity across cohorts[20,32]. Although the definition of stillbirth in the validation cohorts was standardized, stillbirth was defined differently in each model, which prevented head-to-head comparison of model performance.

This study has some limitations. We were able to validate only three of the 40 identified models, mainly due to the failure of studies to adhere to reporting standards for the model equation[13,33]. Only two models were published before the TRIPOD statement. Some cohorts included in the external validation had few observed cases of stillbirth, and only two had more than 100 events. Predicted probabilities in the cohorts only went up to 2%, which makes it difficult for the models to discriminate between women who had and those who did not have the outcome. This further highlights the primary limitation of stillbirth research, which is the comparative rarity of the outcome.

Comparison with existing studies

External validation of prediction models is needed to confirm their generalizability and transportability in populations with different characteristics[34]. However, independent data including a sufficiently large number of stillbirths and relevant predictors for external validation are not readily available. This is one reason why none of the published models has been recommended for use in clinical practice[13]. This meta-analysis demonstrates lower

summary estimates for discrimination than those reported in the development datasets, although this might be due to chance, as some CIs were wide (e.g. for the Smith model); further research is therefore recommended[28–30]. Some published stillbirth prediction models[35,36] have a reported discrimination of >0.8, but the studies either did not report the model equation needed for independent external validation[36] or did not provide sufficient information on predictors[35]. The performance of a prediction model is usually overestimated when estimated only in the dataset used to develop the model, particularly when there are few outcomes relative to the number of predictors considered[37,38]. This study highlights several methodological shortcomings in the development of stillbirth prediction models, which is further reflected in the risk of bias assessment of the models.

Relevance to clinical care

The UK Government and NHS launched a care initiative in a bid to halve the stillbirth rate by 2025, which includes risk assessment as part of a wider care bundle[39]. The bundle does not include tools to help determine if a woman is at increased risk of stillbirth. Instead, individual factors have been identified in order to categorize women as low, moderate or high risk of FGR, which is the most frequent cause of stillbirth in the UK. An accurate tool to predict which women are at increased risk of stillbirth would allow for personalized risk stratification in pregnancy and enable clinicians to make decisions on the need for closer surveillance and timing of delivery in order to prevent fetal death. It would also empower women to make informed decisions based on their risk of stillbirth. This would be a more targeted approach than the currently used system of a generalized population-level risk factor to identify women at risk of stillbirth. However, none of the models validated in this study had sufficient performance or clinical utility to be recommended for use in clinical practice.

Recommendations for further research

Stillbirth prediction models that can be used in routine care would be particularly valuable in low- and middle-income countries, in which the stillbirth burden is disproportionately high. Models which we were unable to validate externally will need to be validated independently before they can be recommended for use. Apart from improvement in the model development process to reduce overfitting by using larger sample sizes and adjusting for optimism of the predictor effects (for example by post-estimation shrinkage or penalizing the model coefficients), additional work is needed to identify novel prognostic factors for use in model development in order to improve the discriminative performance of prediction models[40]. Closer examination of existing stillbirth risk factors could potentially enable inaccurate risk predictors to be abandoned and for clinical care and research to instead be focused on the highest value predictors.

Systematic reviews using aggregate data meta-analysis currently represent the best available evidence on predictors of stillbirth and have proposed several risk factors to categorize women as high risk[41]. However, they

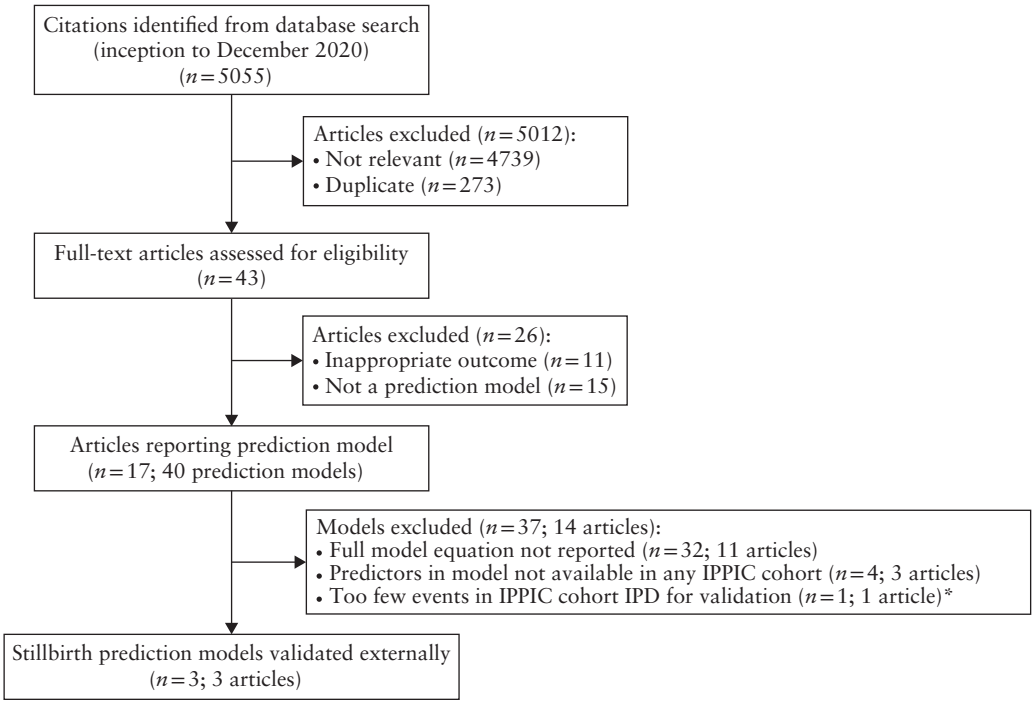

Figure 1 Flow diagram summarizing selection of stillbirth prediction models for external validation in International Prediction of Pregnancy Complications (IPPIC) Network cohorts. *Smith *et al.*[28] reported two models, one of which was validated in this study. IPD, individual participant data.

are limited by heterogeneity among the primary studies, such as in the definition of stillbirth[41]. Existing primary studies are often small, with imprecise estimates and are inconsistent in the confounding factors adjusted for in their analysis, which sometimes leads to contradictory factor-outcome associations. Large cohorts are needed to collect richer data on risk factors in order to enable development and validation of prediction models.

Whilst this study has explored validation of different stillbirth prediction models, stillbirth is the final endpoint of several heterogeneous antecedent pathways, with varying biological mechanisms involved (e.g. those involving FGR and those secondary to diabetes, typically with a large-for-gestational-age infant). It is possible that more than one model will be needed, either for prediction of stillbirth at different gestational ages or for different phenotypes of stillbirth.

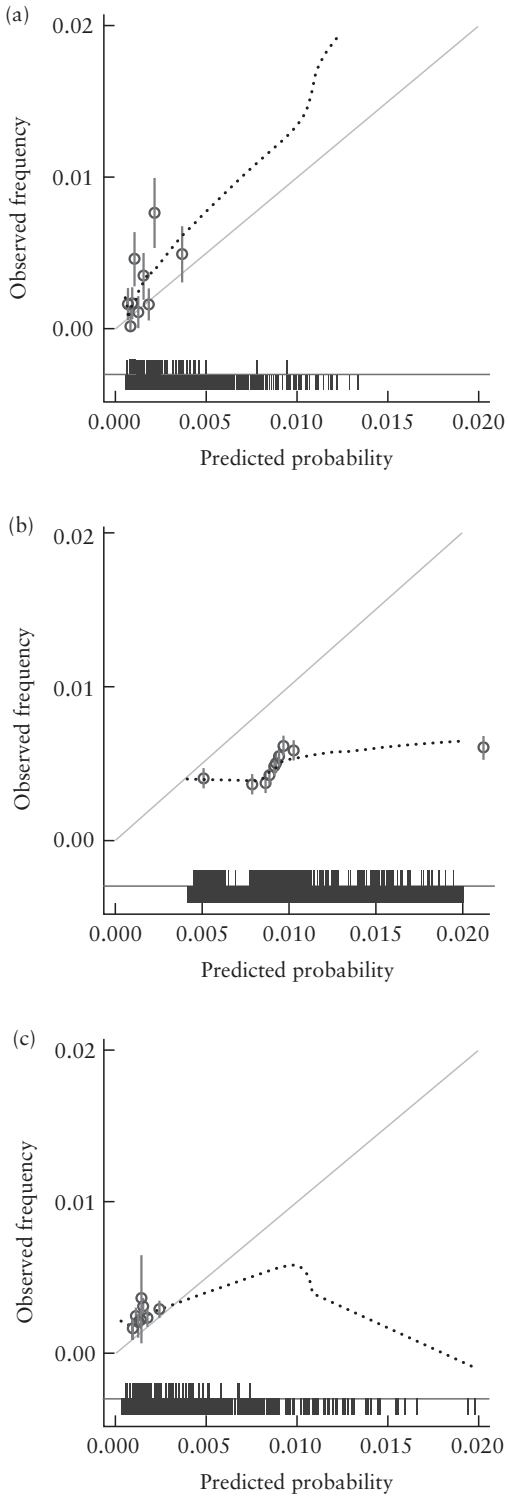

Figure 2 Calibration plots for externally validated stillbirth prediction models, in cohorts with at least 100 events. (a) Smith *et al.*[28] (St George's dataset[42]). (b) Yerlikaya *et al.*[29] (JSOG dataset[47]). (c) Trudell *et al.*[30] (JSOG dataset[47]). Error bars are 95% CI. ——, reference; ·····, locally weighted scatterplot smoother; ——, outcome distribution; ○, risk group.

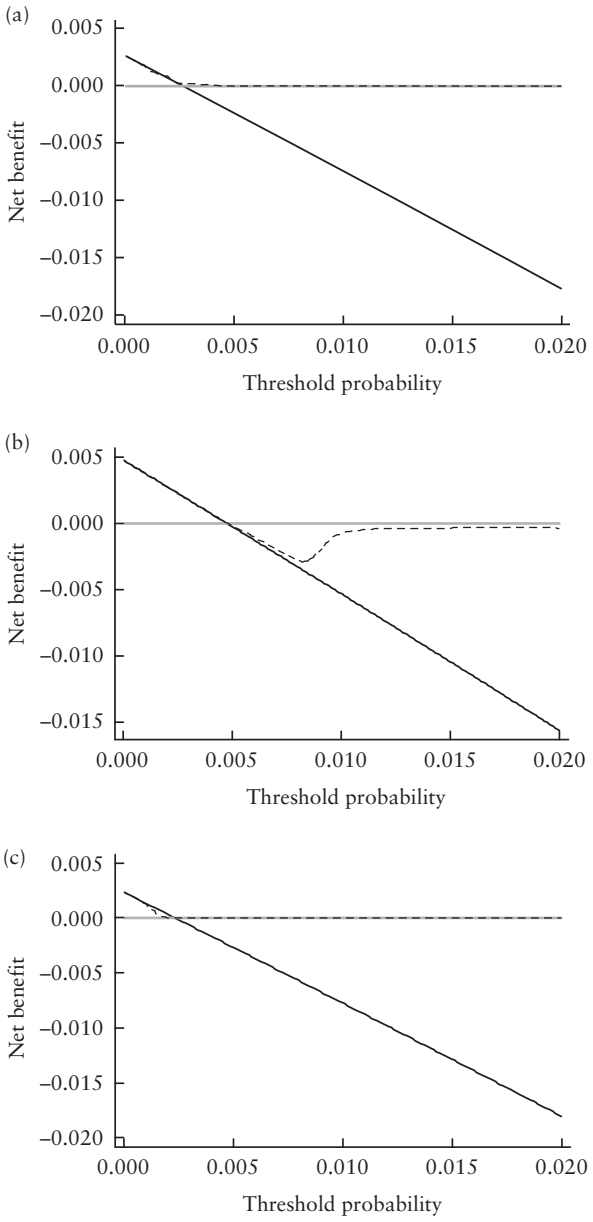

Figure 3 Decision curves for externally validated stillbirth prediction models, in cohorts with at least 100 events. (a) Smith *et al.*[28] (St George's dataset[42]). (b) Yerlikaya *et al.*[29] (JSOG dataset[47]). (c) Trudell *et al.*[30] (JSOG dataset[47]). ——, treat all; ——, treat none; – – –, model.

Conclusions

This study provides a comprehensive assessment and independent external validation of published stillbirth prognostic models across multiple cohorts. The findings suggest methodological shortcomings, including overfitting of models during development. None of the three previously published stillbirth models that were validated in this study showed sufficient performance or clinical utility to be recommended for use in clinical practice. Although there were differences in predictor and outcome definitions used for the different models, all three models considered similar candidate predictors for model development, which may suggest that additional and better predictors (prognostic factors) of stillbirth still need to be identified.

ACKNOWLEDGMENTS

The IPPIC data repository was set up by funding from the National Institute for Health Research Health Technology Assessment Programme (Ref no: 14/158/02). This project was funded by Sands charity. K.I.E.S. is funded by the National Institute for Health Research School for Primary Care Research (NIHR SPCR Launching Fellowship).

The UK Medical Research Council and Wellcome (grant ref: 102215/2/13/2) and the University of Bristol, Bristol, UK, provide core support for ALSPAC. This publication is the work of the authors and J.A., S.T., R.R. and R.W. will serve as guarantors for the contents of this paper.

We acknowledge all researchers who contributed data to this individual participant data meta-analysis, including the original teams involved in the collection of the data and participants who took part in the research studies. We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

REFERENCES

1. Flenady V, Wojcieszek AM, Middleton P, Ellwood D, Erwich JJ, Coory M, Khong TY, Silver RM, Smith GCS, Boyle FM, Lawn JE, Blencowe H, Leisher SH, Gross MM, Horey D, Farrales L, Bloomfield F, McCowan L, Brown SJ, Joseph KS, Zeitlin J, Reinebrant HE, Ravaldi C, Vannacci A, Cassidy J, Cassidy P, Farquhar C, Wallace E, Siassakos D, Heazell AEP, Storey C, Sadler L, Petersen S, Frøen JF, Goldenberg RL. Stillbirths: recall to action in high-income countries. *Lancet* 2016; **387**: 691–702.
2. Flenady V, Koopmans L, Middleton P, Frøen JF, Smith GC, Gibbons K, Coory M, Gordon A, Ellwood D, McIntyre HD, Fretts R, Ezzati M. Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. *Lancet* 2011; **377**: 1331–1340.
3. Draper ES, Gallimore ID, Kurinczuk JJ, Smith PW, Boby T, Smith LK, Manktelow BN, on behalf of the MBRRACE-UK Collaboration. MBRRACE-UK Perinatal Mortality Surveillance Report, UK Perinatal Deaths for Births from January to December 2016. Leicester: The Infant Mortality and Morbidity Studies, Department of Health Sciences, University of Leicester. 2018. https://www.npeu.ox.ac.uk/assets/downloads/mbrrace-uk/reports/MBRRACE-UK%20Perinatal%20Surveillance%20Full%20Report%20for%202016%20-%20June%202018.pdf.
4. Euro-Peristat Project. European Perinatal Health Report. Core indicators of the health and care of pregnant women and babies in Europe in 2015. November 2018. https://www.europeristat.com/images/EPHR2015_Euro-Peristat.pdf.
5. ONS. Vital statistics in the UK: births, deaths and marriages - 2018 update, Office of National Statistics, London, England. https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/birthsummarytablesenglandandwales/2017.
6. Heazell AE, Whitworth MK, Whitcombe J, Glover SW, Bevan C, Brewin J, Calderwood C, Canter A, Jessop F, Johnson G, Martin I, Metcalf L. Research priorities for stillbirth: process overview and results from UK Stillbirth Priority Setting Partnership. *Ultrasound Obstet Gynecol* 2015; **46**: 641–647.
7. Sexton J, Coory M, Kumar S, Smith G, Gordon A, Chambers G, Pereira G, Raynes-Greenow C, Hilder L, Middleton P, Bowman A, Lieske S, Warrilow K, Morris J, Ellwood D, Flenady V. Protocol for the development and validation of a risk prediction model for stillbirths from 35 weeks gestation in Australia. *Diagn Progn Res* 2020; **4**: 21.
8. Townsend R, Manji A, Allotey J, Heazell A, Jorgensen L, Magee LA, Mol BW, Snell K, Riley RD, Sandall J, Smith G, Patel M, Thilaganathan B, von Dadelszen P, Thangaratinam S, Khalil A. Can risk prediction models help us individualise stillbirth prevention? A systematic review and critical appraisal of published risk models. *BJOG* 2021; **128**: 214–224.
9. Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353**: i3140.
10. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG; Cochrane IPD Meta-analysis Methods group. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PLoS Med* 2015; **12**: e1001886.
11. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013; **32**: 3158–3180.
12. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; **68**: 279–289.
13. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015; **162**: 55–63.
14. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, Groupdagger P. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med* 2019; **170**: 51–58.
15. Allotey J, Snell KIE, Chan C, Hooper R, Dodds J, Rogozinska E, Khan KS, Poston L, Kenny L, Myers J, Thilaganathan B, Chappell L, Mol BW, Von Dadelszen P, Ahmed A, Green M, Poon L, Khalil A, Moons KGM, Riley RD, Thangaratinam S; IPPIC Collaborative Network. External validation, update and development of prediction models for pre-eclampsia using an Individual Participant Data (IPD) meta-analysis: the International Prediction of Pregnancy Complication Network (IPPIC pre-eclampsia) protocol. *Diagn Progn Res* 2017; **1**: 16.
16. Snell KIE, Allotey J, Smuk M, Hooper R, Chan C, Ahmed A, Chappell LC, Von Dadelszen P, Green M, Kenny L, Khalil A, Khan KS, Mol BW, Myers J, Poston L, Thilaganathan B, Staff AC, Smith GCS, Ganzevoort W, Laivuori H, Odibo AO, Arenas Ramírez J, Kingdom J, Daskalakis G, Farrar D, Baschat AA, Seed PT, Prefumo F, da Silva Costa F, Groen H, Audibert F, Masse J, Skråstad RB, Salvesen KÅ, Haavaldsen C, Nagata C, Rumbold AR, Heinonen S, Askie LM, Smits LJM, Vinter CA, Magnus P, Eero K, Villa PM, Jenum AK, Andersen LB, Norman JE, Ohkuchi A, Eskild A, Bhattacharya S, McAuliffe FM, Galindo A, Herraiz I, Carbillon L, Klipstein-Grobusch K, Yeo SA, Browne JL, Moons KGM, Riley RD, Thangaratinam S; IPPIC Collaborative Network. External validation of prognostic models predicting pre-eclampsia: individual participant data meta-analysis. *BMC Med* 2020; **18**: 302.
17. Allotey J, Snell KI, Smuk M, Hooper R, Chan CL, Ahmed A, Chappell LC, von Dadelszen P, Dodds J, Green M, Kenny L, Khalil A, Khan KS, Mol BW, Myers J, Poston L, Thilaganathan B, Staff AC, Smith GC, Ganzevoort W, Laivuori H, Odibo AO, Ramírez JA, Kingdom J, Daskalakis G, Farrar D, Baschat AA, Seed PT, Prefumo F, da Silva Costa F, Groen H, Audibert F, Masse J, Skråstad RB, Salvesen KÅ, Haavaldsen C, Nagata C, Rumbold AR, Heinonen S, Askie LM, Smits LJ, Vinter CA, Magnus PM, Eero K, Villa PM, Jenum AK, Andersen LB, Norman JE, Ohkuchi A, Eskild A, Bhattacharya S, McAuliffe FM, Galindo A, Herraiz I, Carbillon L, Klipstein-Grobusch K, Yeo S, Teede HJ, Browne JL, Moons KG, Riley RD, Thangaratinam S. Validation and development of models using clinical, biochemical and ultrasound markers for predicting pre-eclampsia: an individual participant data meta-analysis. *Health Technol Assess* 2020; **24**: 1–252.
18. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res* 2018; **27**: 1634–1649.
19. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med* 2015; **34**: 1841–1863.
20. Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1987. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316696.
21. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; **338**: b605.
22. Hosmer DW, Lemeshow, S. Assessing the Fit of the Model. In *Applied Logistic Regression* (2nd edn). Wiley; New York, NY, 2000; 143–202.
23. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med* 2001; **20**: 3875–3889.
24. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, Viechtbauer W, Simmonds M. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods* 2019; **10**: 83–98.
25. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**: 557–560.
26. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26**: 565–574.

27. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**: i6.
28. Smith GC, Yu CK, Papageorghiou AT, Cacho AM, Nicolaides KH; Fetal Medicine Foundation Second Trimester Screening Group. Maternal uterine artery Doppler flow velocimetry and the risk of stillbirth. *Obstet Gynecol* 2007; **109**: 144–1451.
29. Yerlikaya G, Akolekar R, McPherson K, Syngelaki A, Nicolaides KH. Prediction of stillbirth from maternal demographic and pregnancy characteristics. *Ultrasound Obstet Gynecol* 2016; **48**: 607–612.
30. Trudell AS, Tuuli MG, Colditz GA, Macones GA, Odibo AO. A stillbirth calculator: Development and internal validation of a clinical prediction model to quantify stillbirth risk. *PloS One* 2017; **12**: e0173461.
31. Kleinrouweler CE, Cheong-See Mrcog FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, Mol BW, Pajkrt E, Moons KG, Schuit E. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016; **214**: 79–90.e36.
32. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; **30**: 377–399.
33. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–73.
34. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, Woodward M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**: 691–698.
35. Kayode GA, Grobbee DE, Amoakoh-Coleman M, Adeleke IT, Ansah E, de Groot JA, Klipstein-Grobusch K. Predicting stillbirth in a low resource setting. *BMC Pregnancy Childbirth* 2016; **16**: 274.
36. Aupont JE, Akolekar R, Illian A, Neonakis S, Nicolaides KH. Prediction of stillbirth from placental growth factor at 19–24 weeks. *Ultrasound Obstet Gynecol* 2016; **48**: 631–635.
37. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Moons KGM, Collins G, van Smeden M. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; **368**: m441.
38. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019; **38**: 1276–1296.
39. Saving Babies' Lives Version Two: A care bundle for reducing perinatal mortality. NHS England, 2019. https://www.england.nhs.uk/wp-content/uploads/2019/03/Saving-Babies-Lives-Care-Bundle-Version-Two-Updated-Final-Version.pdf.
40. Riley RD, van der Windt D, Croft P, Moons KGM. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford University Press, Oxford, UK; 2019.
41. Townsend R, Sileo FG, Allotey J, Dodds J, Heazell A, Jorgensen L, Kim VB, Magee L, Mol B, Sandall J, Smith G, Thilaganathan B, von Dadelszen P, Thangaratinam S, Khalil A. Prediction of stillbirth: an umbrella review of evaluation of prognostic variables. *BJOG* 2021; **128**: 238–250.
42. Stirrup OT, Khalil A, D'Antonio F, Thilaganathan B; Southwest Thames Obstetric Research C. Fetal growth reference ranges in twin pregnancy: analysis of the Southwest Thames Obstetric Research Collaborative (STORK) multiple pregnancy cohort. *Ultrasound Obstet Gynecol* 2015; **45**: 301–317.
43. Mone F, Mulcahy C, McParland P, Stanton A, Culliton M, Downey P, McCormack D, Tully E, Dicker P, Breathnach F, Malone FD, McAuliffe FM. An open-label randomized-controlled trial of low dose aspirin with an early screening test for pre-eclampsia and growth restriction (TEST): Trial protocol. *Contemp Clin Trials* 2016; **49**: 143–148.
44. 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**: 2089–2097.
45. Allen RE, Zamora J, Arroyo-Manzano D, Velauthar L, Allotey J, Thangaratinam S, Aquilina J. External validation of preexisting first trimester preeclampsia prediction models. *Eur J Obstet Gynecol Reprod Biol* 2017; **217**: 119–125.
46. Goetzinger KR, Singla A, Gerkowicz S, Dicke JM, Gray DL, Odibo AO. Predicting the risk of pre-eclampsia between 11 and 13 weeks' gestation by combining maternal characteristics and serum analytes, PAPP-A and free beta-hCG. P*renat Diagn* 2010; **30**: 1138–1142.
47. Japan Society of Obstetrics and Gynecology (JSOG). http://www.jsog.or.jp/modules/en/index.php?content_id=1.
48. Jenum AK, Sletner L, Voldner N, Vangen S, Mørkrid K, Andersen LF, Nakstad B, Skrivarhaug T, Rognerud-Jensen OH, Roald B, Birkeland KI. The STORK Groruddalen research programme: A population-based cohort study of gestational diabetes, physical activity, and obesity in pregnancy in a multiethnic population. Rationale, methods, study population, and participation rates. *Scand J Public Health* 2010; **38**(5 Suppl): 60–70.
49. North RA, McCowan LM, Dekker GA, Poston L, Chan EH, Stewart AW, Black MA, Taylor RS, Walker JJ, Baker PN, Kenny LC. Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. *BMJ* 2011; **342**: d1875.
50. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, Ring S, Nelson SM, Lawlor DA. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013; **42**: 97–110.
51. Antsaklis A, Daskalakis G, Tzortzis E, Michalas S. The effect of gestational age and placental location on the prediction of pre-eclampsia by uterine artery Doppler velocimetry in low-risk nulliparous women. *Ultrasound Obstet Gynecol* 2000; **16**: 635–639.
52. Widmer M, Cuesta C, Khan KS, Conde-Agudelo A, Carroli G, Fusey S, Karumanchi SA, Lapaire O, Lumbiganon P, Sequeira E, Zavaleta N, Frusca T, Gülmezoglu AM, Lindheimer MD. Accuracy of angiogenic biomarkers at 20 weeks' gestation in predicting the risk of pre-eclampsia: A WHO multicentre study. *Pregnancy Hypertens* 2015; **5**: 330–338.
53. Andersen LB, Dechend R, Jorgensen JS, Luef BM, Nielsen J, Barington T, Christesen HT. Prediction of preeclampsia with angiogenic biomarkers. Results from the prospective Odense Child Cohort. *Hypertens Pregnancy* 2016; **35**: 405–419.
54. Sibai BM. Management of late preterm and early-term pregnancies complicated by mild gestational hypertension/pre-eclampsia. *Semin Perinatol* 2011; **35**: 292–296.
55. Sibai BM, Caritis SN, Thom E, Klebanoff M, McNellis D, Rocco L, Paul RH, Romero R, Witter F, Rosen M, Depp R; The National Institute of Child Health and Human Development Network of Maternal-Fetal Medicine Units. Prevention of preeclampsia with low-dose aspirin in healthy, nulliparous pregnant women. *N Engl J Med* 1993; **329**: 1213–1218.
56. Holzman C, Bullen B, Fisher R, Paneth N, Reuss L; Prematurity Study Group. Pregnancy outcomes and community health: the POUCH study of preterm delivery. *Paediatr Perinat Epidemiol* 2001; **15** (Suppl 2): 136–158.
57. Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS, Group AS. Vitamins C and E and the risks of preeclampsia and perinatal complications. *N Engl J Med* 2006; **354**: 1796–1806.
58. Savitri AI, Zuithoff P, Browne JL, Amelia D, Baharuddin M, Grobbee DE, Uiterwaal CS. Does pre-pregnancy BMI determine blood pressure during pregnancy? A prospective cohort study. *BMJ Open* 2016; **6**: e011626.
59. van Oostwaard MF, Langenveld J, Bijloo R, Wong KM, Scholten I, Loix S, Hukkelhoven CW, Vergouwe Y, Papatsonis DN, Mol BW, Ganzevoort W. Prediction of recurrence of hypertensive disorders of pregnancy between 34 and 37 weeks of gestation: a retrospective cohort study. *BJOG* 2012; **119**: 840–847.
60. Van Oostwaard MF, Langenveld J, Schuit E, Wigny K, Van Susante H, Beune I, Ramaekers R, Papatsonis DN, Mol BW, Ganzevoort W. Prediction of recurrence of hypertensive disorders of pregnancy in the term period, a retrospective cohort study. *Pregnancy Hypertens* 2014; **4**: 194–202.

SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:

Appendix S1 Members of the IPPIC Collaborative Network

Appendix S2 MEDLINE search strategy for identification of stillbirth prediction models for external validation

Table S1 Studies identified in literature search reporting prediction models for stillbirth

Table S2 PROBAST risk of bias assessment (RoB) of cohorts from the IPPIC Network database used for external validation

Table S3 Proportion of cases with missing (or not recorded) data for each predictor and outcome in each cohort used for external validation

Table S4 Summary of linear predictors and predicted probabilities for each cohort used for external validation