Allotey John (Orcid ID: 0000-0003-4134-6246)
Thilaganathan Basky (Orcid ID: 0000-0002-5531-4301)
Khalil Asma (Orcid ID: 0000-0003-2802-7670)

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

J. Allotey*#[1,2], R. Whittle*[3], K.I.E. Snell[3], M. Smuk[4], R. Townsend[5], P. von Dadelszen[6], A.E.P. Heazell[7], L. Magee[6], G. C.S. Smith[8], J. Sandall[6,9], B. Thilaganathan[5], J. Zamora[1,10,11], R. D. Riley[3], A. Khalil[5], S. Thangaratinam[1,12] for 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 and Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK

[6]Department of Women and Children's Health, School of Life Course Sciences, King's College London, London, UK

[7]Maternal and Fetal Health Research Centre, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, UK

[8] Department of Obstetrics and Gynaecology, NIHR Biomedical Research Centre, Cambridge University, UK

[9]Health Service and Population Research Department, Centre for Implementation Science, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

[10]Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal (IRYCIS), Madrid, Spain

[11]CIBER Epidemiology and Public Health (CIBERESP), Madrid, Spain

[12]Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.23757

*Joint first authors
#Corresponding author: j.allotey.1@bham.ac.uk

Keywords: stillbirth, intra-uterine death, prediction model, individual participant data, external validation

Running Head: Validating stillbirth models using IPD meta-analysis

ABSTRACT

Objective

Stillbirth is a potentially preventable complication of pregnancy. Identifying women at risk can guide decisions on closer surveillance or timing of birth to prevent fetal death. Prognostic models have been developed to predict the risk of stillbirth, but none have yet been externally validated. We externally validated published prediction models for stillbirth using individual participant data (IPD) meta-analysis to assess their predictive performance.

Methods

We searched Medline, EMBASE, DH-DATA and AMED databases from inception to December 2020 to identify stillbirth prediction models. We included studies that developed or updated prediction models for stillbirth for use at any time during pregnancy. IPD from cohorts within the International Prediction of Pregnancy Complication (IPPIC) Network were used to externally validate the identified prediction models whose individual variables were available in the IPD. We assessed the risk of bias of the models and IPD using PROBAST, and reported discriminative performance using the C-statistic, and calibration performance using calibration plots, calibration slope and calibration-in-the-large. We estimated performance measures separately in each study, and then summarised across studies using random-effects meta-analysis. Clinical utility was assessed using net benefit.

Results

We identified 17 studies reporting the development of 40 prognostic models for stillbirth. None of the models were previously externally validated, and only a fifth (20%, 8/40) reported the full model equation. We were able to validate three of these models using the IPD from 19 cohort studies (491,201 pregnant women) within the IPPIC Network database. Based on evaluating their development studies, all three models had an overall high risk of bias according to PROBAST. In our IPD meta-analysis, the models had summary C-statistics ranging from 0.53 to 0.65; summary calibration slopes of 0.40 to 0.88, and generally with observed risks predictions that were too extreme compared

to observed risks; and little to no clinical utility as assessed by net benefit. However, there remained uncertainty in performance for some models due to small available sample sizes

Conclusion

The three validated models generally showed poor and uncertain predictive performance in new data, with limited evidence to support their clinical application. Findings suggest methodological shortcomings in their development including overfitting of models. Further research is needed to further validate these and other models, identify stronger prognostic factors, and to develop more robust prediction models.

INTRODUCTION

Stillbirth continues to be a major burden globally, accounting for almost two thirds of perinatal mortality.[1,2] In the UK, stillbirth rates were largely unchanged from 2000 – 2015, and at 4.2 stillbirths/1,000 births in 2017 had one of the highest rates in Europe.[3-5] Prediction and individualisation of risk remain key priorities for stillbirth research,[6,7] because accurate identification of women at risk of stillbirth can guide decisions on closer surveillance, or timing of birth to prevent fetal death. A recent review that identified existing prediction models for stillbirth reported that none had been externally validated.[8] As a result, no prediction models are routinely used in clinical practice and none have been recommended by any national or international guidelines.

An 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 devasting outcome limits rigorous investigation of existing stillbirth prediction models in single cohort studies. An individual participant data (IPD) meta-analysis that combines the raw data from multiple studies, has great potential for use in externally validating 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 also allows us to evaluate the generalisability 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, critically appraise and externally validate existing multivariable prognostic models for stillbirth prediction using IPD meta-analysis within the independent International Prediction of Pregnancy Complication (IPPIC) Network database, and to assess the clinical utility of the models using decision curve analysis.

METHODS

This study was based on a prospective protocol registered on PROSPERO (registration number CRD42018074788), and reported in line with TRIPOD recommendations for reporting risk prediction model validation studies.[13]

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

We systematically searched Medline, EMBASE, DH-DATA and AMED databases 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 hand searched reference lists of relevant articles and systematic reviews to identify potentially eligible studies. Our search included terms for stillbirth, intrauterine fetal death and perinatal mortality, and study selection was done independently by two researchers. The complete search strategy is provided in appendix 1.

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

We included 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 reported the model equation in the publication. No attempts were made to contact authors of studies that did not publish their model equation. Given the wide international variation in definitions of stillbirth, we accepted the authors' definition of stillbirth (both antepartum and intrapartum), and included models developed for use at any time in pregnancy. We excluded models that: predicted stillbirth as part of a composite adverse outcome; contained predictors that were not measured in any of the cohorts within the IPPIC IPD; or if there were too few outcomes (<10 stillbirths) reported across the IPPIC IPD cohorts with the same predictors as the model, to allow for its external validation.

We extracted data on the definition of stillbirth, number of participants and events, population type, predictors in the final model, and the reported model performance. Based on information in the original articles, 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, and this was

done independently by two researchers. Disagreement were resolved through discussions with a third researcher. We classified the risk of bias to be low, high or unclear for each domain, as well as an overall risk of bias. Each domain included signalling questions rated as “yes”, “probably yes”, “probably no”, “no” or “no information”. Domains with any signalling 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 it scored low in all domains, high if any one domain had a high risk of bias, and unclear for any other classifications.

International Prediction of Pregnancy Complications (IPPIC) Network

We identified cohorts for the IPPIC Network by systematically reviewing evidence for risk of pregnancy complications including pre-eclampsia, stillbirth and fetal growth restriction (FGR), and inviting research groups that had undertaken the primary studies to join the IPPIC Network and share their primary IPD. We also searched major databases and repositories and contacted researchers within the IPPIC Network to identify relevant studies or datasets that may have been missed, including unpublished research and birth cohorts. We formatted, cleaned and harmonised datasets received and assessed the quality of each cohort using the participants, predictors and outcome domains of the PROBAST tool.[14] Study population could vary from low to high risk of development of complications. The network includes nearly 150 collaborators from 26 countries, contributing IPD of 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 regularly being enriched with additional studies. We consider the predictor variables contained within the IPPIC Network 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 harmonised have previously been published.[15-17]

Statistical analysis for external validation using IPPIC network database

Data harmonisation and set-up

Predictors or outcomes of existing prediction models that were partially missing for <95% of individuals in any cohort were multiply imputed under the missing at random assumption using multiple imputation by chained equations.[18,19] We used linear regression to impute for approximately normally distributed continuous variables, logistic regression for binary variables, and multinomial logistic regression for categorical variables. We carried out multiple imputation for each individual cohort separately and generated fifty imputed datasets for each. We also included other predictors that were available within the cohort as auxiliary variables in the imputation models. Imputation checks were completed by looking at 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 were summarised 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 (SE) for each performance statistic in each cohort.[20]

The discriminatory performance of models were assessed using the C-statistic (summarised as the area under receiver operating characteristic curve, where 1 indicates perfect discrimination and 0.5 indicates no discrimination beyond chance), and calibration statistics of the 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 that model predictions are systematically too low or too high across the cohort, ideal value of 0).[21 22] Model calibration was also visually assessed using calibration plots representing the average predicted probability for risk groups categorised using deciles of predicted probability against the observed proportion in each group, in cohorts with at least 100 events. A lowess smoother curve was applied to show calibration across the entire range of predicted probabilities at the individual-

level (i.e. without categorisation). 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 summarised using a random effects meta-analysis to calculate a summary estimate for the model's discrimination and calibration performance. Model performance was summarised for each statistic as the average and 95% confidence interval (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 summarised. We also reported the approximate 95% prediction intervals, for potential predictive performance in a new study, as calculated using the approach of Higgins et al.[26]

Decision curve analysis

We performed decision curve analysis (DCA) to assess the clinical value of the models on 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 generally very low risk of stillbirth), compared to either simply classifying all women as having the outcome or no women as having the outcome.[27] The strategy with the highest net benefit at a particular threshold has the highest clinical value.[28] The net benefit is represented as a function of the decision threshold in decision curve plots.

All statistical analyses were performed using Stata software version 15.

RESULTS

From 5055 citations we identified 17 articles describing the development of 40 stillbirth prediction models published between 2007 and 2020 (Appendix 2). Three studies reporting three prediction models - Smith 2007,[29] Yerlikaya 2016,[30] and Trudell 2017[31] met our inclusion criteria for external validation in the IPPIC IPD datasets (Figure 1).

Characteristics of included models

All three models were developed using binary logistic regression in unselected populations of pregnant women,[29-31] and the definition of stillbirth varied between the studies. Two models included only maternal clinical characteristics as predictors,[30,31] while one model additionally included ultrasound markers.[29] Only one study had at least 10 events per predictor for model development,[30] 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. The characteristics of included studies and models are described in Table 1.

Characteristics of the IPPIC validation cohorts

Of the 78 cohorts in the IPPIC data repository, 19 cohorts (24%) contained relevant data that could be used to externally validate at least one of the three prediction models identified. Only women with singleton pregnancies in the cohorts were used for external validation. The prevalence of stillbirth ≥24 weeks gestation in the cohorts ranged from 0.1% - 1.6%. A quarter of the studies used for external validation included only low risk (26%, 5/19) women, while a fifth (21%, 4/19) included only high-risk women in the cohorts. Seventy-five percent (14/19) of the cohorts used for external validation had an overall low risk of bias as assessed by PROBAST, 21% (4/19) were assessed as high risk and one cohort as unclear (appendix 3). Summary maternal characteristics and outcomes of women in the validation cohort are provided in table 2, and a summary of missing data for each predictor and outcome is provided in appendix 4.

External validation and meta-analysis of predictive performance

The Smith 2007 model[29] was validated in 3 cohorts, Yerlikaya 2016 model[30] in 4 cohorts and the Trudell 2017 model[31] in 17 cohorts. Two of the cohorts used to validate the Smith 2007 model and all four of the cohorts used to validate the Yerlikaya 2016 model were

also used to validate the Trudell 2017 model. A direct comparison of performance of the prediction models was not possible due to differences in outcomes of each model. The distribution of the linear predictor and predicted probability for each model and validation cohort are shown in appendix 5.

Model predictive performance

The C-statistics of models in the different validation cohorts ranged from 0.56-0.82 in the Smith 2007 model, 0.54-0.73 in the Yerlikaya 2016 model and 0.34-0.69 in the Trudell 2017 model (Table 3). The Trudell 2017 model had the lowest overall discrimination across the validation cohorts. Summary C-statistics of the models were 0.65 (95% CI 0.53 to 0.75) for the Smith 2007 model, 0.61 (95% CI 0.43 to 0.77) for the Yerlikaya 2016 model, and 0.53 (95% CI 0.51 to 0.55) for the Trudell 2017 model (Table 4). Confidence intervals for the Smith 2007 and Yerlikaya 2016 models were wide, due to the fewer 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 of overfitting during model development; in particular, the 95% confidence intervals for the calibration slope were all below 1 for the Yerlikaya 2016 and Trudell 2017 models, indicating extreme predictions compared to what was observed (Table 4).

Each of the three models were validated in one cohort with at least 100 events. The average calibration plots showed miscalibration of the predicted risk of stillbirth in all three models (Figure 2). However, 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 2007 model, due to less data on stillbirth outcome in the validation cohorts available for this model, and so further research is needed for this model.

Net benefit of model use

The 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 a fifth of published stillbirth prognostic models reported the model equation required for independent external validation. Three models developed in high-income countries could be externally validated using cohorts from the IPPIC data repository. The models were mostly developed using maternal clinical characteristics, but one model additionally included ultrasound markers. PROBAST of the original model development articles suggested risk of bias concerns, and our IPD meta-analysis of model performance showed low discriminatory ability and poor calibration, with calibration slopes mostly <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 at least one cohort with >100 events, confidence intervals of predictive performance were wide for the Smith 2007 model, suggesting 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.[8,32] Our 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 through an IPD meta-analysis. We used multiple imputation of predictors and outcomes for each cohort separately, to avoid loss of useful information, and ensure we did not mask any heterogeneity across cohorts.[20,33] Although the definition of stillbirth in the validation cohorts were standardised, stillbirth was defined differently in each model, which prevented a head-to-head comparison of model performance.

Our study has some limitations. We were only able to validate three of the 40 identified models , mainly due to the failure of studies to adhere to reporting standards of publishing the model equation.[34,35] Only two models were published before release of TRIPOD. Some cohorts used in the external validation had few observed cases of stillbirths, and only two had more than 100 events. Predicted probabilities in the cohorts only went up to 3%, which makes it difficult for the models to discriminate between women who had and did not have the outcome. This further highlights the primary limitation of stillbirth research, which is the comparative rarity of the outcome.

Comparison to existing studies

External validation of prediction models are needed to confirm generalisability and transportability of a model in populations with different characteristics.[36] However, independent data with sufficiently large sample sizes of stillbirth and relevant predictors for external validation of models are not readily available. This is a factor on why none of the published models have been recommended for use in clinical practice.[35] Our meta-analysis obtained lower summary estimates for discrimination to that reported in the development datasets, although this might be due to chance as some confidence intervals were wide (e.g. Smith 2007), further research is recommended.[29-31] Some published stillbirth models report discrimination of > 0.8,[37,38] but these studies either did not report the model equation needed for independent external validation,[38] or did not provide enough information on predictors .[37] In most cases, the performance of a prediction model is often overestimated when only estimated in the dataset used to develop the model, especially when there are few outcomes relative to the number of predictors considered.[39,40] Our study highlighted 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 stillbirth rates by 2025, which includes risk assessment as part of a wider care-bundle.[41] The bundle does not include tools to help determine if a woman is at increased risk of stillbirth, instead individual factors have been identified to categorise women as low, moderate or high risk of FGR, the most frequent cause of stillbirth in the UK. An accurate tool to predict which woman is at increased risk of stillbirth would allow for personalised risk stratification in pregnancy, and enable clinicians to make decisions on closer surveillance, or timing of birth to prevent fetal death. It would also empower mothers to make informed decisions on their risk of stillbirth. This would be a more targeted approach than the currently used system of a generalised 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 practice.

Recommendations for further research

Stillbirth prediction models that can be used in routine care would be especially valuable in low-and-middle-income countries, where stillbirth burden is disproportionately high. Models we were unable to externally validate will need to be independently validated 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 (e.g. by post-estimation shrinkage or penalising the model coefficients), additional work is needed to identify novel prognostic factors for use in model development, to improve the discriminatory performance of prediction models.[42] A closer examination of existing stillbirth risk factors could potentially enable us to abandon inaccurate risk predictors and focus clinical care and research 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 categorise women as high-risk.[43] However, these studies are limited by heterogeneity in the data reported within the primary studies, such as in the definition of stillbirth.[43] Existing primary studies are often small with imprecise estimates, and inconsistencies in 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 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 (for example, 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 at different gestational ages, or for stillbirths with similar phenotypes.

CONCLUSION

This is a comprehensive assessment and independent external validation of published stillbirth prognostic models across multiple cohorts. 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 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 additional and better predictors (prognostic factors) of stillbirth still need to be identified.

Abbreviations

IPD	Individual participant data
IPPIC	International Prediction of Pregnancy Complications
PROBAST	Prediction study Risk of Bias Assessment
SE	Standard error
CI	Confidence interval
LP	Linear predictor

Declarations

Ethics approval and consent to participate

Not applicable. The study involved secondary analysis of existing anonymised data.

Consent for publication

Not applicable

Availability of data and materials

The data that support the findings of this study are available from the IPPIC data sharing committee, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of contributing collaborators.

Competing interests

None to declare

Study registration

PROSPERO ID: CRD42018074788

Funding

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. Kym Snell 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 provide core support for ALSPAC. This publication is the work of the authors and JA, ST, RR, and RW will serve as guarantors for the contents of this paper.

Authors' contributions

ST, AK developed the protocol. RW wrote the statistical analysis plan and performed the analysis, JA produced the first draft of the article and revised the article. RR and KS

oversaw the statistical analyses and analysis plan. MS and JA formatted, harmonised and cleaned IPPIC datasets, in preparation for analysis. JA, MS mapped the variables in the datasets, and cleaned and quality checked the data. JA, ST, MS and RT undertook the literature searches, study selection, acquired Individual Participant Data, contributed to the development of all versions of the manuscript and led the project. All authors provided input at all stages of the project and helped revise the article.

Acknowledgements

The following are members of the IPPIC Collaborative Network[+]

Arri Coomarasamy - University of Birmingham; Alex Kwong - University of Bristol; Ary I. Savitri - University Medical Center Utrecht; Kjell Åsmund Salvesen - Norwegian University of Science and Technology; Sohinee Bhattacharya - University of Aberdeen; Cuno S.P.M. Uiterwaal - University Medical Center Utrecht; Annetine C. Staff - University of Oslo; Louise Bjoerkholt Andersen - University of Southern Denmark; Elisa Llurba Olive - Hospital Universitari Vall d'Hebron; Christopher Redman - University of Oxford; Line Sletner - University of Oslo; George Daskalakis - University of Athens; Maureen Macleod - University of Dundee; Baskaran Thilaganathan - St George's University of London; Mali Abdollahain - RMIT University; Javier Arenas Ramírez - University Hospital de Cabueñes; Jacques Massé - Laval University; Asma Khalil - St George's University of London; Francois Audibert - Université de Montréal; Per Minor Magnus - Norwegian Institute of Public Health; Anne Karen Jenum - University of Oslo; Ahmet Baschat - Johns Hopkins University School of Medicine; Akihide Ohkuchi - University School of Medicine, Shimotsuke-shi; Fionnuala M. McAuliffe - University College Dublin; Jane West - University of Bristol; Lisa M. Askie - University of Sydney; Fionnuala Mone - University College Dublin; Diane Farrar - Bradford Teaching Hospitals; Peter A. Zimmerman-Päijät-Häme Central Hospital; Luc J.M. Smits - Maastricht University Medical Centre; Catherine Riddell - Better Outcomes Registry & Network (BORN); John C. Kingdom - University of Toronto; Joris van de Post - Academisch Medisch Centrum; Sebastián E. Illanes - University of the Andes; Claudia Holzman - Michigan State University; Sander M.J. van Kuijk - Maastricht University Medical Centre; Lionel Carbillon - Assistance Publique-Hôpitaux de Paris Université; Pia M. Villa - University of Helsinki and Helsinki University Hospital; Anne Eskild - University of Oslo; Lucy Chappell - King's College

London; Federico Prefumo - University of Brescia; Luxmi Velauthar – Queen Mary University of London; Paul Seed - King's College London; Miriam van Oostwaard - IJsselland Hospital; Stefan Verlohren - Charité University Medicine; Lucilla Poston - King's College London; Enrico Ferrazzi - University of Milan; Christina A. Vinter - University of Southern Denmark; Chie Nagata - National Center for Child Health and Development, Tokyo, Japan; Mark Brown - University of New South Wales; Karlijn C. Vollebregt - Academisch Medisch Centrum; Satoru Takeda - Juntendo University, Tokyo, Japan; Josje Langenveld - Atrium Medisch Centrum Parkstad; Mariana Widmer - World Health Organization; Shigeru Saito - University of Toyama, Toyama, Japan; Camilla Haavaldsen - Akershus University Hospital; Guillermo Carroli - Centro Rosarino De Estudios Perinatales; Jørn Olsen - Aarhus University; Hans Wolf - Academisch Medisch Centrum; Nelly Zavaleta - Instituto Nacional De Salud; Inge Eisensee - Aarhus University; Patrizia Vergani - University of Milano-Bicocca; Pisake Lumbiganon - Khon Kaen University; Maria Makrides - South Australian Health and Medical Research Institute; Fabio Facchinetti - Università degli Studi di Modena e Reggio Emilia; Evan Sequeira - ga Khan University; Robert Gibson - University of Adelaide; Sergio Ferrazzani - Università Cattolica del Sacro Cuore; Tiziana Frusca - Università degli Studi di Parma; Jane E. Norman - University of Bristol; Ernesto A. Figueiró-Filho - Mount Sinai Hospital; Olav Lapaire - Universitätsspital Basel; Hannele Laivuori - University of Helsinki and Helsinki University Hospital; Jacob A. Lykke – Rigshospitalet; Agustin Conde-Agudelo - Eunice Kennedy Shriver National Institute of Child Health and Human Development; Alberto Galindo - Universidad Complutense de Madrid; Alfred Mbah - University of South Florida; Ana Pilar Betran - World Health Organisation; Ignacio Herraiz - Universidad Complutense de Madrid; Lill Trogstad - Norwegian Institute of Public Health; Gordon G.S. Smith - Cambridge University; Eric A.P. Steegers - University Hospital Nijmegen; Read Salim - HaEmek Medical Center; Tianhua Huang - North York General Hospital; Annemarijne Adank - Erasmus Medical Centre; Jun Zhang - National Institute of Child Health and Human Development; Wendy S. Meschino - North York General Hospital; Joyce L Browne - University Medical Centre Utrecht; Rebecca E. Allen - Queen Mary University of London; Fabricio Da Silva Costa - University of São Paulo; Kerstin Klipstein-Grobusch Browne - University Medical Centre Utrecht; Caroline A. Crowther - University of Adelaide; Jan Stener Jørgensen - Syddansk Universitet; Jean-Claude Forest - Centre hospitalier universitaire de Québec; Alice R. Rumbold - University of Adelaide; Ben W. Mol - Monash University; Yves Giguère - Laval University; Louise C. Kenny - University

of Liverpool; Wessel Ganzevoort - Academisch Medisch Centrum; Anthony O. Odibo - University of South Florida; Jenny Myers - University of Manchester; SeonAe Yeo - University of North Carolina at Chapel Hill; Francois Goffinet - Assistance publique – Hôpitaux de Paris; Lesley McCowan - University of Auckland; Eva Pajkrt - Academisch Medisch Centrum; Helena J. Teede - Monash University and Monash Health; Bassam G. Haddad - Portland State University; Gustaaf Dekker - University of Adelaide; Emily C. Kleinrouweler - Academisch Medisch Centrum; Édouard LeCarpentier - Centre Hospitalier Intercommunal Creteil; Claire T. Roberts - University of Adelaide; Henk Groen - University Medical Center Groningen; Ragnhild Bergene Skråstad - St Olavs Hospital; Seppo Heinonen - University of Helsinki and Helsinki University Hospital; Kajantie Eero - University of Helsinki and Helsinki University Hospital; Dewi Anggraini - University of Lambung Mangkurat; Athena Souka - University of Athens Medical School; Jose Guilherme Cecatti - University of Campinas; Ilza Monterio - University of Campinas; Athanasios Pillalis - University of Athens; Renato Souza - University of Campinas; Lee Ann Hawkins - University of Calgary; Rinat Gabbay- Benziv - Hillel Yaffe Medical Center; Francesca Crovetto - University of Barcelona; Francesc Figuera - University of Barcelona, Laura Jorgensen - Queen Mary University of London, Julie Dodds - Queen Mary University of London, Mehali Patel - Sands, stillbirth and neonatal death charity, London, Amir Aviram - University of Toronto, Toronto, Ontario, Canada, Aris Papageorghiou - St George's University of London, London, UK, Khalid Khan - University of Granada, Granada, Spain

We would like to acknowledge all researchers who contributed data to this IPD 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, et al. Stillbirths: recall to action in high-income countries. *The Lancet* 2016; **387**(10019): 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. *The Lancet* 2011; **377**(9774): 1331-40.

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. .

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. Available www.europeristat.com.

5. ONS (2018) Vital statistics in the UK: births, deaths and marriages - 2018 update, Office of National Statistics, London, England https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigrati..

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 in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology* 2015; **46**(6): 641-7.

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 SN, 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, 10 March 2020, PREPRINT (Version 1) available at Research Square [+https://doi.org/10.21203/rs.3.rs-16494/v1+]. 2020.

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 : an international journal of obstetrics and gynaecology* 2020.

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 IPDM-aMg. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PLoS medicine* 2015; **12**(10): 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. *Statistics in Medicine* 2013; **32**(18): 3158-80.

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. *Journal of clinical epidemiology* 2015; **68**(3): 279-89.

13. Collins GS, Reitsma JB, Altman DG, Moons KG, for the members of the Tg. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *European urology* 2014.

14. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S; PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of internal medicine* 2019; **170**(1): 51-8.

15. Allotey J, Snell KIE, Chan C, Hooper, R., Dodds, J., Rogozinska, E., Khan, K. S., Poston, L., Kenny, L., Myers, J., Thilaganathan, B., Chappell, L., Mol, B. W., Von Dadelszen, P., Ahmed, A., Green, M., Poon, L., Khalil, A., Moons, K. G. M., Riley, R. D. and Thangaratinam, S. 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, et al. for the IPPIC Collaborative Network. External validation of prognostic models predicting pre-eclampsia: Individual participant data meta-analysis. BMC Medicine 2020 (in press)

17. Allotey J, Snell KIE, Smuk M, et al, for the IPPIC Collaborative Network. Accuracy of clinical characteristics, biochemical and ultrasound markers in predicting pre-eclampsia: External validation and development of prediction models using an Individual Participant Data (IPD) meta-analysis Health Technol Assess 2020 (in press).

18. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res* 2016.

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. *Statistics in medicine* 2015; **34**(11): 1841-63.

20. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley; 1987.

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. Applied Logistic Regression. 2nd ed. New York: Wiley; 2000: 143-202.

23. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in medicine* 2001; **20**(24): 3875-89.

24. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Res Synth Methods* 2019; **10**(1): 83-98.

25. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003; **327**(7414): 557-60.

26. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society Series A, (Statistics in Society)* 2009; **172**(1): 137-59.

27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26**(6): 565-74.

28. 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.

29. Smith GC, Yu CK, Papageorghiou AT, Cacho AM, Nicolaides KH, Fetal Medicine Foundation Second Trimester Screening G. Maternal uterine artery Doppler flow velocimetry and the risk of stillbirth. *Obstet Gynecol* 2007; **109**(1): 144-51.

30. Yerlikaya G, Akolekar R, McPherson K, Syngelaki A, Nicolaides KH. Prediction of stillbirth from maternal demographic and pregnancy characteristics. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology* 2016; **48**(5): 607-12.

31. 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**(3): e0173461.

32. Kleinrouweler CE, Cheong-See Mrcog FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *American journal of obstetrics and gynecology* 2015.

33. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in medicine* 2011; **30**(4): 377-99.

34. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* 2015; **162**(1): W1-73.

35. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Annals of internal medicine* 2015; **162**(1): 55-63.

36. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**(9): 691-8.

37. Kayode GA, Grobbee DE, Amoakoh-Coleman M, et al. Predicting stillbirth in a low resource setting. *BMC pregnancy and childbirth* 2016; **16**: 274.

38. Aupont JE, Akolekar R, Illian A, Neonakis S, Nicolaides KH. Prediction of stillbirth from placental growth factor at 19-24 weeks. *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology* 2016; **48**(5): 631-5.

39. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020; **368**: m441.

40. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in medicine* 2019; **38**(7): 1276-96.

41. Saving Babies' Lives Version Two: A care bundle for reducing perinatal mortality https://www.england.nhs.uk/wp-content/uploads/2019/03/Saving-Babies-Lives-Care-Bundle-Version-Two-Updated-Final-Version.pdf Accessed 15th October 2020.

42. Riley RD, van der Windt D, Croft P, Moons KGM. Prognosis Research in Healthcare: Concepts, Methods and Impact. Oxford, UK: Oxford University Press; 2019.

43. Townsend R, Sileo FG, Allotey J, et al. Prediction of stillbirth: an umbrella review of evaluation of prognostic variables. *BJOG : an international journal of obstetrics and gynaecology* 2020.

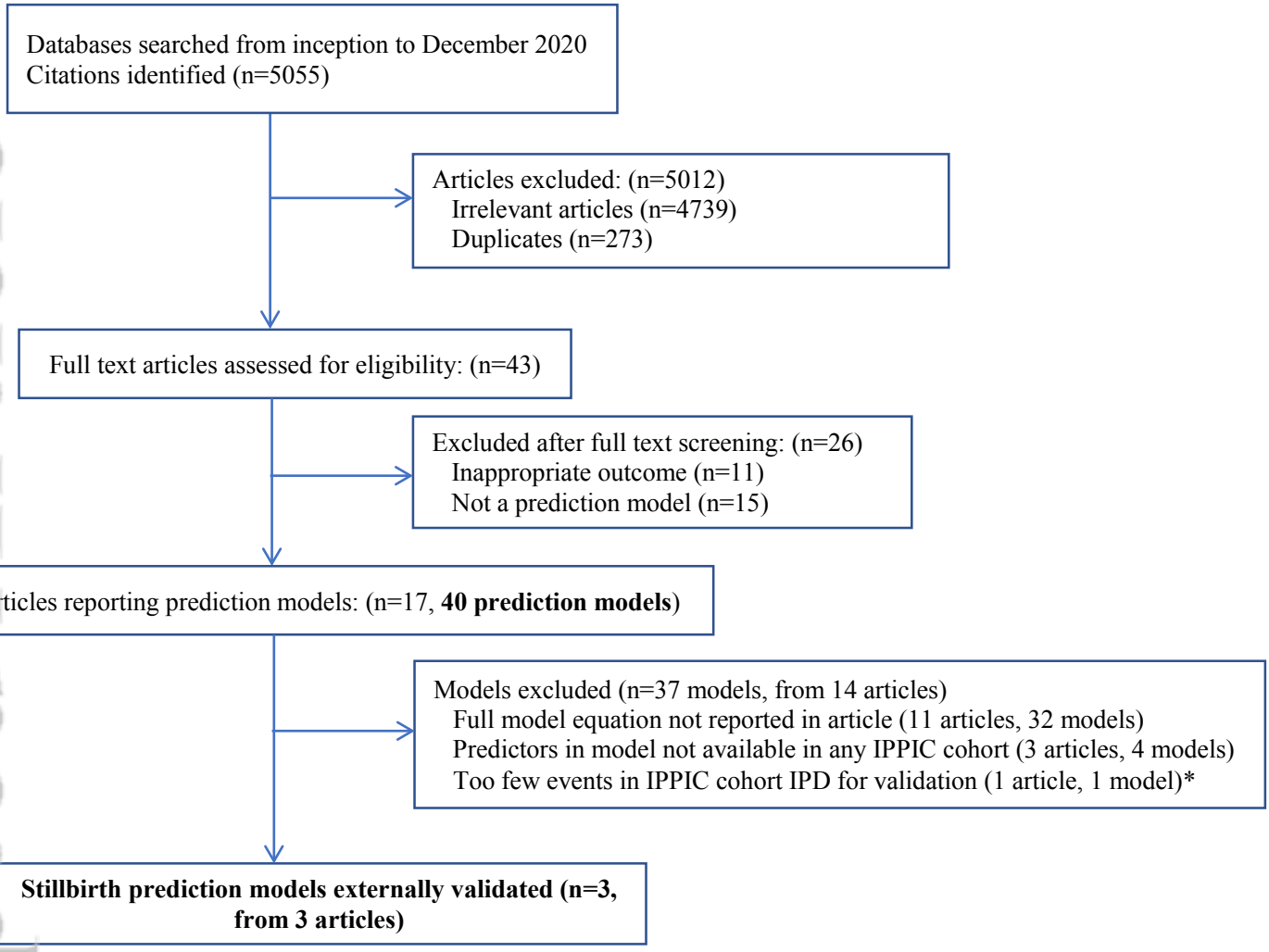
FIGURE LEGENDS

Figure 1: Flow diagram of prediction models identified for external validation in IPPIC cohorts

Figure 2: Calibration plots for externally validated stillbirth prediction models in cohorts with greater than 100 events

Figure 3: Decision curves for externally validated stillbirth prediction models in cohorts with greater than 100 events

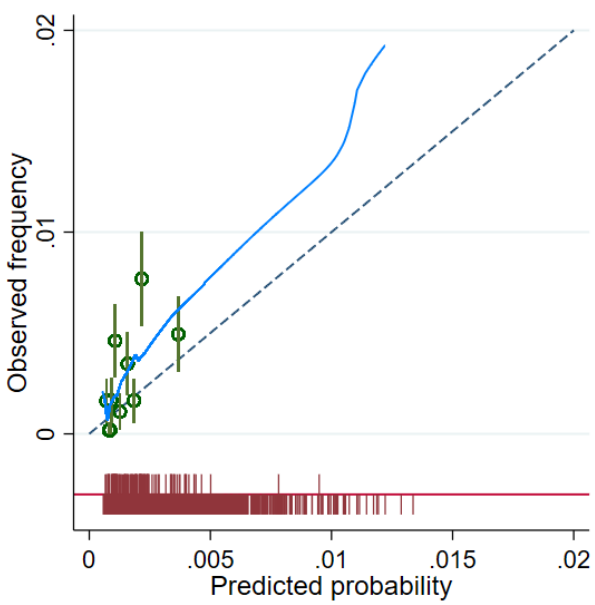
Figure 1: Flow diagram of prediction models identified for external validation in IPPIC cohorts



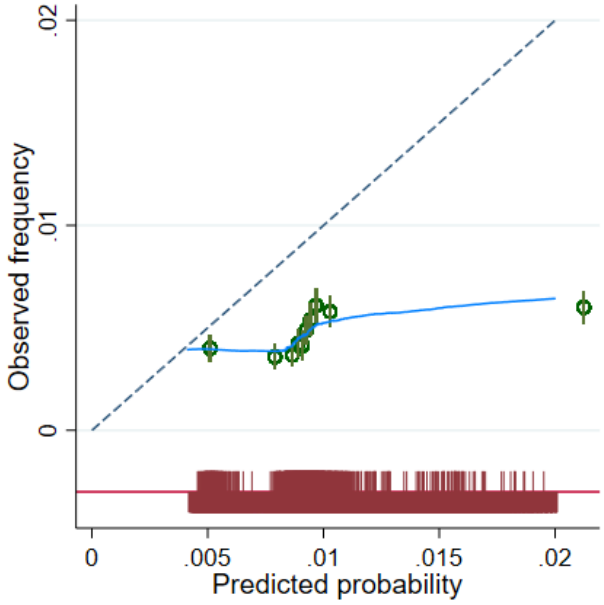
**Smith et al Second Trimester Screening G. Maternal uterine artery Doppler flow velocimetry and the risk of stillbirth. Obstet Gynecol 2007; 109(1): 144-51 reported two models, one of which was validated in this study*

Figure 2: Calibration plots for externally validated stillbirth prediction models in cohorts with greater than 100 events

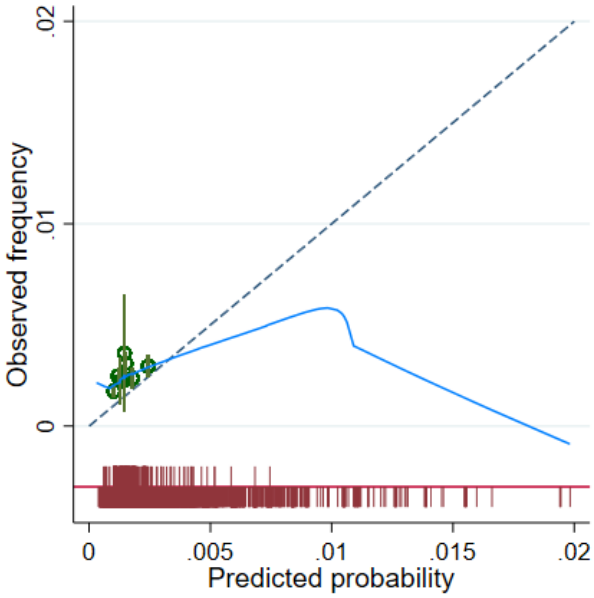
Smith 2007 model in the St Georges dataset



Yerlikaya 2016 model in the JSOG dataset



Trudell 2017 model in the JSOG dataset



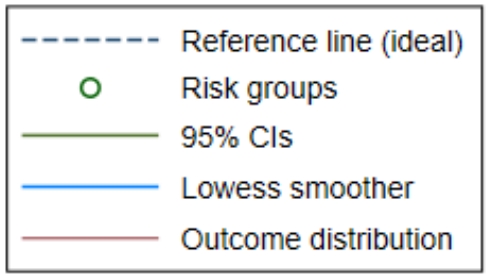
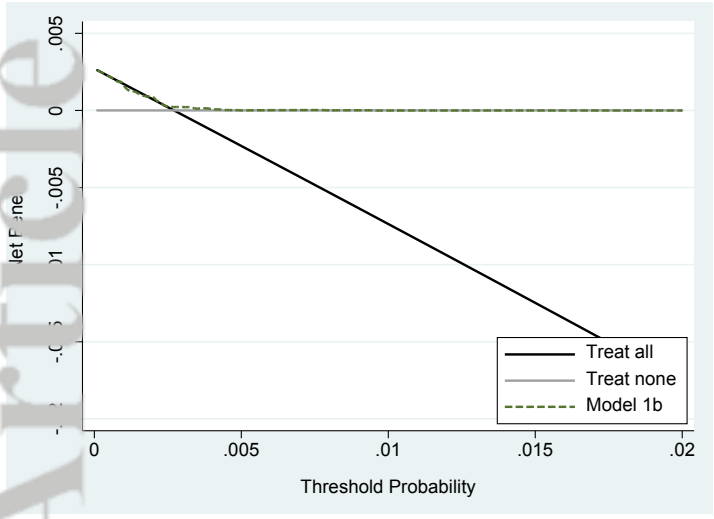
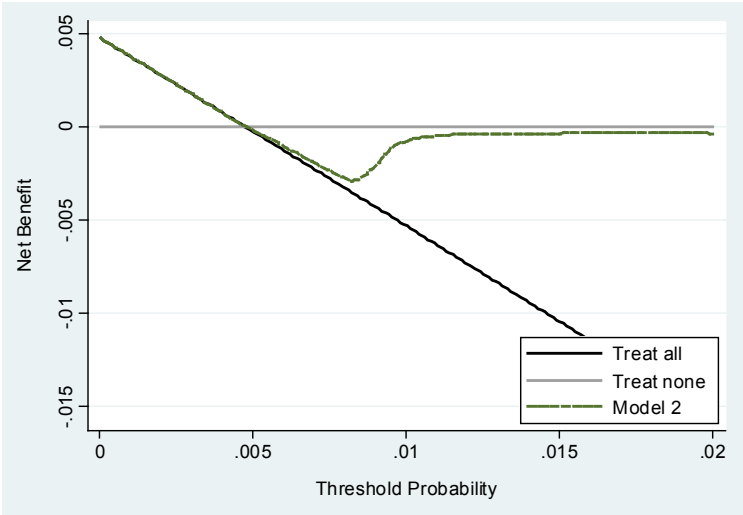


Figure 3: Decision curves for externally validated stillbirth prediction models in cohorts with greater than 100 events

Smith 2007 model in the St Georges dataset



Yerlikaya 2016 model in the JSOG dataset



Trudell 2017 model in the JSOG dataset

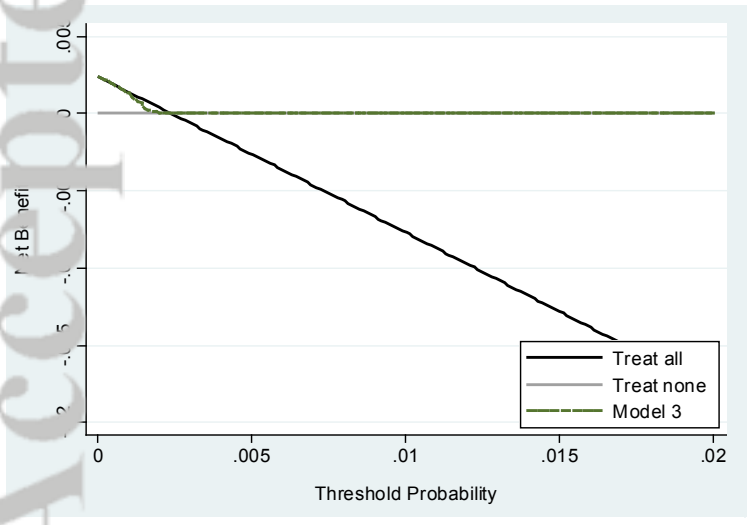


Table 1: Characteristics of studies and prediction models included in the external validation.

Author, year, Country	Population	No. of women	No. of candidate predictors	Predictors included in model	Prediction model equation for linear predictor (LP)*	Outcome; Gestation at stillbirth	No. of events	Discrimination AUC (95% CI)	PROBAST RoB of model
Smith 2007, UK	Pregnant women between 22-24 weeks gestation, excluding those with short cervix from 7 hospitals	30,519	17	Uterine artery pulsatility index, BMI, Ethnicity	LP = - 7.806 + 0.867(mean pulsatility index) + 0.768(if BMI 25-29.9) + 0.768(if BMI≥30) + 0.624(if African-American ethnicity)	Stillbirth ≥33 weeks	109	0.67 (0.60-0.75)	High
Yerlikaya 2016, UK	Women with singleton pregnancies between 11-25 weeks gestation, attending 2 hospitals for routine pregnancy care	113,415	17	Weight, Ethnicity, Assisted conception, Smoking, Hypertension, APS, SLE, Diabetes, Previous Stillbirth	LP = - 6.02615 + 0.01037(weight(kg) – 69) + 0.70027(if Afro-Caribbean ethnicity) + 0.57994(if assisted conception) + 0.53367(if smoke cigarettes) + 0.96253(if chronic hypertension) + 1.28416(if APS or SLE) + 0.93628(if diabetic) + 1.57086(if parous with previous stillbirth)	Stillbirth ≥24 weeks	396	0.64 (0.61-0.67)	High
Trudell 2017, USA	Women with singleton pregnancies in their second trimester, attending for routine anatomic screening	57,326	NR	Maternal age, Ethnicity, Parity, BMI, Smoking, Hypertension, Diabetes	LP = - 6.8772 – 0.8707(if maternal age < 18) + 0.2094(if maternal age 35-39) + 0.4377(if maternal age > 40) + 0.8536(if black race) + 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 pre-gestational diabetes)	Stillbirth ≥32 weeks	330	0.66 (0.60-0.72)	High

BMI=body mass index; APS=antiphospholipid syndrome; SLE=systemic lupus erythematosus, RoB=Risk of Bias; AUC=Area Under the Curve; CI=Confidence Interval

* For logistic regression, logit(p)=LP where the linear predictor (LP) = $\alpha + \beta_1 * x_1 + \beta_2 * x_2 + \ldots$, and absolute predicted probabilities (p) can be obtained using the transformation $p = \frac{e^{LP}}{1+e^{LP}}$.

Table 2: Summary maternal characteristics and outcomes of IPPIC individual participant data used for external validation

Dataset	No. of pregnancies	Population type	Maternal age: mean (SD); range	BMI: median [IQR], range	White ethnicity, n (%)	Nulliparous, n (%)	Outcome, n (%)		
							≥24weeks	≥32weeks	≥33weeks
St Georges	54635	Any pregnancy	30.5 (5.6); 13 to 54	23.5 [21.3, 26.8]; 13 to 54	33257 (62)	29313 (54)	233 (0.43)	160 (0.29)	148 (0.27)
Test	557	Low risk	32.0 (4.8); 18 to 43	24 [21.6, 27.1]; 17.4 to 45.2	539 (97)	557 (100)	5 (0.92)	4 (0.73)	4 (0.73)
POP	4212	Any pregnancy	29.9 (5.1); 16 to 48	24.1 [21.8, 27.3]; 14.7 to 54.7	3,900 (93)	4212 (100)	11 (0.26)	8 (0.19)	8 (0.19)
Allen	1045	Any pregnancy	29.9 (5.1); 15 to 48	23.6 [21.0, 26.8]; 14.8 to 51.1	398 (38)	584 (56)	3 (0.29)	3 (0.29)	3 (0.29)
Goetzinger	4035	Any pregnancy	34.8 (4.4); 16 to 52	24.4 [21.8, 28.8]; 15.4 to 62.4	3282 (83)	751 (20)	15 (0.37)	15 (0.37)	15 (0.37)
JSOG	379390	Any pregnancy	32.2 (5.4); 10 to 59	20.5 [19.0, 22.6]; 10.5 to 69.8	0 (0)	195983 (52)	1792 (0.47)	895 (0.24)	801 (0.21)
StorkG	812	Any pregnancy	29.8 (4.8); 19 to 45	25.1 [22.3, 28.4]; 16.2 to 49.8	375 (46)	377 (46)	6 (0.74)	5 (0.62)	4 (0.49)
SCOPE	5628	Low risk	28.7 (5.5); 14 to 45	24.2 [21.9, 27.5]; 15.4 to 58.5	5061 (90)	5628 (100)	17 (0.30)	9 (0.16)	8 (0.14)
ALSPAC	15038	Any pregnancy	27.7 (4.9); 13 to 46	21.5 [19.7, 23.7]; 11.7 to 61.3	11769 (97)	5704 (45)	41 (0.27)	27 (0.18)	26 (0.17)
Antsaklis	3328	Low risk	30.9 (4.8); 14 to 47	22.7 [20.6, 25.7]; 14.5 to 50.1	3229 (97)	3328 (100)	2 (0.06)	2 (0.06)	2 (0.06)
WHO	7273	High risk	22.5 (5.8); 11 to 51	23.1 [21.0, 26.1]; 13.5 to 54.8	2222 (31)	6710 (92)	8 (0.46)	8 (0.46)	8 (0.46)
Andersen	2120	Any pregnancy	30.2 (4.5); 17 to 45	23.4 [21.2, 26.2]; 14.9 to 49.9	1765 (97)	1193 (56)	6 (0.28)	4 (0.19)	4 (0.19)
NICHD HR	1848	High risk	27.1 (6.3); 15 to 43	28.4 [23.5, 35.0]; 13.4 to 68.5	612 (33)	430 (23)	23 (1.26)	8 (0.44)	8 (0.44)
NICHD LR	3097	Low risk	20.6 (4.4); 15 to 39	22.7 [20.4, 25.7]; 13.4 to 51.2	548 (18)	3097 (100)	13 (0.44)	6 (0.20)	6 (0.20)
POUCH	3019	Any pregnancy	26.4 (5.8); 15 to 47	27.7 [24.3, 32.9]; 15.1 to 66.3	2018 (67)	1293 (43)	10 (0.33)	4 (0.13)	4 (0.13)

Rumbold	1877	Low risk	26.4 (5.7); 13 to 44	24.1 [21.5, 27.6]; 13.7 to 57.6	1777 (95)	1877 (100)	11 (0.59)	9 (0.48)	9 (0.48)
Indonesian cohort	2223	Any pregnancy	28.6 (5.9); 10 to 59	22.9 [20.1, 26.3]; 13.3 to 67.6	0 (0)	664 (43)	12 (0.70)	6 (0.35)	6 (0.35)
Van Oostwaard 2012	425	High risk	32.0 (4.1); 23 to 42	24.3 [21.5, 27.9]; 16.2 to 41.8	288 (84)	0 (0)	2 (1.05)	2 (1.05)	2 (1.05)
Van Oostwaard 2014	639	High risk	32.1 (4.4); 21 to 43	25.9 [22.5, 31.2]; 17.7 to 56.5	360 (72)	0 (0)	5 (1.64)	3 (0.98)	3 (0.98)

Table 3: Study specific performance statistics

Author, year	Outcome	Study	N Total	No. Events (%)	Performance statistic (95% CI)		
					C-statistic	Calibration slope	Calibration-in-the-large
Smith 2007	Stillbirth ≥33 weeks	St Georges	54635	148 (0.27)	0.65 (0.60 to 0.70)	0.87 (0.57 to 1.16)	0.57 (0.41 to 0.73)
		TEST	557	4 (0.72)	0.82 (0.52 to 0.95)	1.57 (0.16 to 2.99)	1.74 (0.75 to 2.72)
		POP	4212	8 (0.18)	0.56 (0.36 to 0.75)	0.49 (-0.93 to 1.92)	0.29 (-0.41 to 0.98)
Yerlikaya 2016	Stillbirth ≥24 weeks	Allen	1045	3 (0.29)	0.64 (0.31 to 0.88)	0.54 (-1.57 to 2.65)	-1.52 (-2.66 to -0.39)
		Goetzinger	4035	26 (0.64)	0.63 (0.42 to 0.80)	0.66 (-0.10 to 1.42)	-1.98 (-2.37 to -1.59)
		JSOG	379390	1802 (0.47)	0.54 (0.53 to 0.56)	0.44 (0.32 to 0.55)	-0.74 (-0.79 to -0.70)
		StorkG	812	7 (0.86)	0.73 (0.56 to 0.85)	1.04 (-0.42 to 2.50)	-0.41 (-1.15 to 0.34)
Trudell 2017	Stillbirth ≥32 weeks	Scope	5628	9 (0.16)	0.34 (0.20 to 0.51)	-1.84 (-3.77 to 0.86)	-0.03 (-0.69 to 0.62)
		Allen	1045	3 (0.29)	0.47 (0.18 to 0.79)	-0.28 (-3.43 to 2.87)	0.58 (-0.56 to 1.71)
		ALSPAC	15038	27 (0.18)	0.48 (0.33 to 0.63)	-0.04 (-1.77 to 1.68)	0.15 (-0.23 to 0.53)
		Goetzinger	4035	24 (0.59)	0.54 (0.27 to 0.79)	0.52 (-0.70 to 1.75)	1.20 (0.78 to 1.62)
		Antsaklis	3328	2 (0.06)	0.43 (0.10 to 0.84)	-1.08 (-4.72 to 2.57)	-1.27 (-2.64 to 0.10)
		WHO	7273	63 (0.87)	0.54 (0.40 to 0.67)	0.17 (-0.73 to 1.07)	1.73 (1.00 to 2.46)
		Andersen	2120	4 (0.19)	0.62 (0.28 to 0.87)	1.55 (-2.00 to 5.10)	0.25 (-0.73 to 1.23)
		NICHD HR	1848	8 (0.43)	0.61 (0.39 to 0.80)	0.44 (-0.57 to 1.44)	-0.03 (-0.72 to 0.67)
		NICHD LR	3097	7 (0.23)	0.64 (0.35 to 0.85)	0.88 (-0.60 to 2.36)	0.05 (-0.76 to 0.85)
		POUCH	3019	4 (0.13)	0.64 (0.42 to 0.81)	0.66 (-1.10 to 2.42)	-0.38 (-1.36 to 0.60)
		Rumbold	1877	9 (0.48)	0.47 (0.27 to 0.69)	-0.68 (-2.64 to 1.28)	1.07 (0.42 to 1.73)
		JSOG	379390	897 (0.24)	0.53 (0.51 to 0.55)	0.41 (0.18 to 0.65)	0.49 (0.43 to 0.56)
		Indonesian cohort	2223	11 (0.49)	0.69 (0.48 to 0.85)	1.92 (0.07 to 3.78)	1.30 (0.57 to 2.02)
		StorkG	812	6 (0.74)	0.43 (0.16 to 0.76)	0.29 (-1.79 to 2.37)	1.58 (0.77 to 2.39)
		Van Oostwaard 2012	425	14 (3.29)	0.64 (0.35 to 0.86)	0.65 (-0.75 to 2.05)	2.89 (1.71 to 4.06)
		Van Oostwaard 2014	639	4 (0.63)	0.59 (0.24 to 0.87)	0.38 (-1.48 to 2.24)	1.20 (0.03 to 2.37)
		POP	4212	8 (0.19)	0.63 (0.40 to 0.82)	1.20 (-0.42 to 2.81)	0.09 (-0.61 to 0.78)

Table 4: Summary estimates of performance statistics from meta-analysis

Author, year	Outcome	No. of validation cohorts	Total events	Total pregnancies	Summary estimate of performance statistic (95% CI), Measures of heterogeneity (I^2, τ^2)		
					C-statistic	Calibration slope	Calibration-in-the-large
Smith 2007	≥33 weeks	3	160	59404	0.65 (0.53 to 0.75) I^2=0%, τ^2=0	0.88 (0.26 to 1.50) I^2=0%, τ^2=0	0.76 (-0.95 to 2.48) I^2=76.6%, τ^2=0.292
Yerlikaya 2016	≥24 weeks	4	1838	385282	0.61 (0.43 to 0.77) I^2=48.6%, τ^2=0.102	0.45 (0.26 to 0.63) I^2=0%, τ^2=0	-1.15 (-2.35 to 0.05) I^2=91.4%, τ^2=0.462
Trudell 2017	≥32 weeks	17	1100	436009	0.53 (0.51 to 0.55) I^2=0%, τ^2=0	0.40 (0.19 to 0.62) I^2=0%, τ^2=0	0.64 (0.18 to 1.11) I^2=89.1%, τ^2=0.552