DR. ROSEMARY TOWNSEND (Orcid ID : 0000-0002-3438-7069)
MR. JOHN ALLOTEY (Orcid ID : 0000-0003-4134-6246)
PROF. ALEXANDER HEAZELL (Orcid ID : 0000-0002-4303-7845)
BEN WJ MOL (Orcid ID : 0000-0001-8337-550X)
GORDON SMITH (Orcid ID : 0000-0003-2124-0997)
DR. PETER VON DADELSZEN (Orcid ID : 0000-0003-4136-3070)

Article type : Systematic review

Can risk prediction models help us individualise stillbirth prevention? A systematic review and critical appraisal of published risk models

Townsend R[1,2], Manji A[2], Allotey J[3,4], Heazell AEP[5,6], Jorgensen L[7], Magee LA[8], Mol BW[9], Snell KIE,[10] Riley RD[10], Sandall J[11], Smith GCS[12], Patel M[13], Thilaganathan B[1,2], von Dadelszen P[8], Thangaratinam S[3,4], Khalil A[1,2].

1. Molecular and Clinical Sciences Research Institute, St George's, University of London and St George's University Hospitals NHS Foundation Trust, London, UK
2. Fetal Medicine Unit, St George's University Hospitals NHS Foundation Trust, London, UK
3. Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK
4. Pragmatic Clinical Trials Unit, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK
5. St. Mary's Hospital, Manchester Academic Health Science Centre, Manchester University NHS Foundation Trust, Manchester, UK.
6. Maternal and Fetal Health Research Centre, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK.
7. Katie's Team, East London, United Kingdom.

8. School of Life Course Sciences, Faculty of Life Sciences and Medicine, King's College London, London, United Kingdom.
9. Department of Obstetrics and Gynaecology, School of Medicine, Monash University, Melbourne, Australia
10. Centre for Prognosis Research, School of Primary, Community and Social Care, Keele University.
11. Department of Women and Children's Health, School of Life Course Sciences, Faculty of Life Sciences & Medicine, King's College London, St. Thomas' Hospital, London, United Kingdom
12. Department of Obstetrics and Gynaecology, University of Cambridge, NIHR Cambridge Biomedical Research Centre, Cambridge, United Kingdom
13. Sands (Stillbirth and Neonatal Death Society), London, UK

Corresponding: Professor Asma Khalil

Fetal Medicine Unit
Lanesborough Wing
St George's NHS Foundation Trust
Blackshaw Road
Tooting
LONDON
SW17 0QT

akhalil@sgul.ac.uk
+447917400164

Running head: Systematic review of stillbirth risk prediction tools

ABSTRACT

Background: Stillbirth prevention is an international priority - risk prediction models could individualise care and reduce unnecessary intervention, but their use requires evaluation.

Objectives: To identify risk prediction models for stillbirth, and assess their potential accuracy and clinical benefit in practice.

Search strategy: Medline, EMBASE, DH-DATA and AMED databases were searched from inception to June 2019 using terms relevant to stillbirth, perinatal mortality and prediction models. The search was compliant with PRISMA guidelines.

Selection criteria: Studies developing and/or validating prediction models for risk of stillbirth developed for application during pregnancy.

Data collection and analysis: Study screening and data extraction were conducted in duplicate, using the CHARMS checklist. Risk of bias was appraised using the PROBAST tool.

Results: The search identified 2751 citations. Fourteen studies reporting development of 69 models were included. Variables consistently included were: ethnicity, body mass index (BMI), uterine artery Doppler, pregnancy-associated plasma protein (PAPP-A) and placental growth factor (PlGF). Almost all models had significant concern about risk of bias. Apparent model performance (i.e. in the development dataset) was highest in models developed for use later in pregnancy and including maternal characteristics, and ultrasound and biochemical variables, but few were internally validated and none were externally validated.

Conclusions: Almost all models identified were at high risk of bias. There are first trimester models of possible clinical benefit in early risk stratification; these require validation and clinical evaluation. There were few later pregnancy models, but if validated, these could be most relevant to individualised discussions around timing of birth.

Funding

The authors are collaborators in the IPPIC (International Prediction of Pregnancy Complications) stillbirth project, funded by Sands (the Stillbirth and Neonatal Death Society).

Keywords: stillbirth, prediction, model, epidemiology, perinatal, Systematic reviews, Fetal medicine, serum screening

Tweetable abstract: *Prediction models using maternal factors, blood tests and ultrasound could individualise stillbirth prevention, but existing models are at high risk of bias.*

INTRODUCTION

There is substantial patient and clinician interest in individualising obstetric care, and risk prediction models are proliferating.(1) Stillbirth accounts for more global deaths than HIV/AIDS or cancer; over 2.6 million a year. (2) As reduction of stillbirth has become an international health policy priority, induction of labour rates have increased (3). Therefore, accurate risk stratification and individualisation of interventions for the prevention of stillbirth are a research priority in order to minimise iatrogenic harm and facilitate effective stillbirth prevention.

As clinicians increasingly apply prediction models in practice, critical appraisal of model quality and clinical impact is crucial. Guidelines for robust model development and reporting exist to support best practice. (4,5) (6) In 2016, a systematic review reported on over 100 prognostic models developed for use in obstetrics.(1) Few were either internally or externally validated. Internal validation is crucial to estimate and, if necessary, adjust for optimism in apparent model performance in the development dataset. External validation evaluates predictive performance in a new independent dataset. Often a model provides less accurate predictions in a new population, and therefore continuous validation and updating of models for clinical use is necessary, as is systematic evaluation of clinical impact.(7)

The aim of this review was to identify studies reporting on the development and/or validation of models for the prediction of stillbirth (intrauterine fetal death after 20 weeks' gestation, encompassing all international definitions of stillbirth(8)) during pregnancy and assess their methodological quality and potential for further external validation and/or clinical use.

METHODS

This review was conducted according to guidance from the Cochrane Prognosis Methods Group, and utilising the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) guidance.(4,9) The findings were reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA).(10) This systematic review was prospectively registered with the PROSPERO database (Ref: CRD42018074788). Patients were not directly involved in the conduct of this study. The authors are collaborators in the IPPIC (International Prediction of Pregnancy Complications) stillbirth project, funded by Sands (the Stillbirth and Neonatal Death Society).

We included models intended for use by maternity care providers at any time during pregnancy in either high or low resource settings. The expected aim of such models would be to select women for interventions, additional monitoring or scheduled birth (by induction of labour or planned Caesarean section) to prevent stillbirth.

Electronic searches were made of the Medline, EMBASE, Allied & Complementary Medicine and DH-DATA databases from inception to June 2019. The search included relevant terms for stillbirth, intrauterine fetal death and perinatal mortality combined with terms to increase sensitivity and specificity for prediction models, unrestricted by language.(11,12) (Appendix S1) Reference lists of included studies and studies citing existing systematic reviews of stillbirth prediction identified in the search were reviewed in order to identify additional potentially relevant papers to be included in abstract screening.

We defined a prediction model as a model, score or clinical decision tool incorporating more than three variables to estimate the patient-specific risk of stillbirth. We defined variables included in model development as candidate predictors, and variables included in the final model as predictors. We included development and validation studies of any prediction model addressing the risk of stillbirth at any time in during pregnancy. We accepted and noted the authors' definition of stillbirth because this varies between settings.(8) We excluded studies that assessed the first trimester screening 'combined test' as prediction tool for stillbirth, since this model was not developed for this purpose. We excluded studies exploring prediction of composite outcomes even where the composite included stillbirth, unless model performance for stillbirth alone was reported.

Abstracts were screened, potentially eligible texts were retrieved and examined, and data extracted in duplicate (RT, AM). Data were extracted according to the CHARMS checklist.(4) Discrepancies were resolved by consensus.

Every model reported was assessed using PROBAST criteria (5) (see Appendix S2) and an overall assessment of 'high' or 'low' risk of bias was made by consensus of both reviewers. PROBAST includes domains for model participants, predictors, outcomes and analysis. For the latter, PROBAST considers the analytical techniques used during model development to determine the risk

of model bias and subsequent underperformance in a new population. Model performance may be apparent (performance in the development dataset without adjustment for overfitting), validated internally (e.g. via bootstrap, cross or split sample validation) or externally (in an independent data set). Performance measures include both discrimination (ability of the model to separate those who will develop the outcome of interest from those who will not) and calibration (difference between predicted and observed risks across the population and the whole range of predicted risk). (13)

A key variable in model development is the sample size, particularly the number of events per predictor parameter (EPP). The commonly held 'rule of thumb' is that >10 EPP minimises the risk of model overfitting.(14) However, the optimal sample size is actually context specific and may be higher or lower than 10,(15,16) taking into account outcome prevalence, the magnitude of predictor effects and the expected fit of the model (R^2). (17–19)

For the analysis, where multiple models were presented, the authors' final recommended model was included. The results are presented as counts and percentages as indicated. If multiple validation reports of a single model were identified, we planned to undertake meta-analysis of model performance, but no such reports were identified.

Extraction of performance statistics

For each model we extracted data on performance statistics including the c statistic or AUC (Area Under the receiver operating characteristic Curve), calibration in the large (CITL), and sensitivity and specificity at particular risk thresholds. We recorded the presentation of calibration plots and extracted calibration slope, mean absolute error and 'goodness-of-fit' where reported. Where these measures were not reported directly we did not seek to derive them indirectly from other information.

RESULTS

The literature search identified 2751 studies. Fifty-seven were selected for full text screening and 14 papers (published 2007-2018) reporting 69 models were included. (Figure 1, Figure S1) The characteristics of the included studies are described in Table S1. No external validation studies were identified. Where development of multiple models was reported in one paper, this was because different groups of candidate predictors were included or the outcome predicted was varied. Model

developers varied candidate predictors because of their availability in the setting of intended use (20,21) or to investigate the contribution of novel predictors to model performance. (22,23)

Three groups developed models using data derived from low-resource settings (20,21,24) while the remainder were developed in higher-resource settings. Most included all women presenting for routine pregnancy care. Some excluded women for whom delivery information was unavailable. (23,25–28) Several models were developed for high-risk populations – women admitted to hospital,(20) women with hypertension in pregnancy,(21) high BMI (29) or women requiring third trimester fetal ultrasound.(30) One was developed using a low risk population excluding women with previous adverse obstetric outcomes, 'infection' and medical co-morbidities.(31)

The predictors included in the final models are summarised in Table 1. In some studies the full set of candidate predictors was not clear.(24,32) The most frequently included predictor was ethnicity. Ethnicity was identified as a predictor of stillbirth in both univariable and multivariable analyses in every study that evaluated it, but was highly variable in classification, even among datasets from the same country and city. (Table S2) One group divided their population by country of birth (29), one by regionally specific ethnic groups (20) who would all have been classified as a single ethnicity in other models. One US based group classified as 'Black, White or other' while UK groups included one or more categories for Asian women. (25,31,33) Although ethnicity is likely to intersect with social disadvantage as a risk factor for stillbirth(34), only two studies(20,29) included measures of social disadvantage as candidate predictors. In one, occupation and rural residence were included with ethnicity in the final model as predictors.(20)

Maternal body mass index (BMI) and uterine artery pulsatility index (UtAPI) were also consistently included in prediction models when evaluated as candidate predictors. Three reports excluded ultrasound candidate predictors because they were not routinely available.(20,21,32) Other maternal characteristics included as predictors in the included models were smoking and alcohol use, maternal education, prior pregnancy loss, parity and place of residence. One model included 'maternal medical co-morbidities' as a predictor (20) while others included individual conditions including diabetes, systemic lupus erythematosus (SLE), and hypertension.(25–27,32)

Where biomarkers were evaluated, pregnancy-associated plasma protein A (PAPP-A) and placental growth factor (PlGF) each consistently contributed to prediction. Where they were compared directly, PlGF made a greater contribution to performance.(27)

All identified models were developed for the prediction of stillbirth, but stillbirth was variably defined. No core outcome set (COS) for stillbirth research has yet been published, and researchers relied on national, international or customised outcome specifications. In all, 16 distinct stillbirth outcomes were reported across the 69 models (Table S3). The most common gestational cut off was >24 weeks (range 20-34). Most studies excluded pregnancies affected by congenital anomaly and several excluded women who delivered spontaneously <24 weeks from the development data set. Some groups subclassified stillbirth by gestation (<32, <33, >33, <37 and >37 weeks of gestation) or categorised stillbirth by cause (unexplained or placentally associated). All identified models used either antepartum or 'all stillbirth'; none predicted intrapartum stillbirth.

Quality assessment

The risk of bias and applicability of each of the models are reported in detail in Table S1 and summarised in Figures 2a and 2b.

Description of clinical context and population was of a high standard. For some models there was concern about bias related to the unclear exclusion and inclusion criteria. In most, definition and measurement of candidate predictors and outcomes was acceptable. All included studies were retrospective and none included predictors masked from the clinical teams or outcome assessors, but knowledge of predictors would be unlikely to change determination of the outcome in this context. One study did not specify the candidate predictors used in model development.(32)

Almost all included models raised significant concern about risk of bias relating to conduct and reporting of the analysis. According to the PROBAST criteria, an EPP <10 flags potential for concern, whilst an EPP >20 indicates less potential for concern.(5) Only eight were developed with an EPP of >20, chiefly those predicting a broader outcome like 'all stillbirth' or 'all antepartum stillbirth'. This EPP is based on a generalisation and for a rare outcome like stillbirth, fewer EPP may be sufficient. For example, if stillbirth prevalence were 1% (and in many contexts it is lower), then around 4 EPP could be adequate to minimise overfitting.(14–17,35) Twelve models had an EPP <4 which raises

large concern about overfitting, even acknowledging the low prevalence of stillbirth. These included all models predicting stillbirth >37 weeks. Other concerns in analysis were inappropriately categorised predictors; BMI, age and UtAPI were all frequently categorised when they could have been continuous. Where continuous variables were used, few reported assessment for non-linearity. Missing data were handled by complete case analysis, with only one study using multiple imputation.(20)

Most studies reported model discrimination, but only three included a description of model calibration. (20,21,32) Kayode presented calibration plots and reported the mean absolute error, Payne presented calibration plots and tested 'goodness-of-fit' using the Hosmer-Lemeshow method while Trudell explored calibration in terms of centiles of probability but did not present calibration plots or formal assessment of calibration. Four studies used internal validation incorporating bootstrapping to assess for optimism. (20,21,29,32) Two studies updated the models based on their findings. No other studies described internal validation.

Nine reports gave the model equation; overall 15 models were reported with the intercept and coefficients. A further three were made available for use on a web portal but the algorithms were not provided.(23,25,27)

The only models with an overall low risk of bias (20) were developed for use in low resource country settings using only clinical information, and, as might be expected, apparent model performance was lower than in models including ultrasound and biomarkers. Moreover, this model is likely to require recalibration to be generalisable.

Model performance

Model performance was most frequently described using AUC. The best performing model (maternal characteristics, ultrasound and PlGF in the second trimester to predict placentally associated stillbirth <32 weeks) reported excellent apparent discrimination with AUC 0.990 (0.983-0.998).(23) This model was at high risk of bias because it used an effective sample size of 90 events, an EPP of 6 and was not assessed or adjusted for overfitting.

Direct comparison of model performance was limited by the fact that each were developed in different datasets with different populations and contexts. In general, second and third trimester models had better apparent discrimination than earlier models. Models incorporating biomarkers and ultrasound findings had higher performance than maternal characteristics alone. (Figure S2) One model incorporating solely ultrasound variables (estimated fetal weight [EFW], cerebroplacental ratio [CPR] and femur length [FL]) in the third trimester in a high risk population had a reported AUC of 0.88 (0.77-0.99) in the development dataset, superior to many incorporating maternal characteristics with or without biomarkers, although it was also at high risk of bias. (30) Although discrimination was higher the more specific the outcome chosen
("placentally-associated stillbirth <32 weeks" rather than "all stillbirth") these models were also likely to be limited by small sample sizes and low EPP. Figure 3 shows the AUC of models predicting a) "all stillbirth" and b) "stillbirth >37 weeks". No studies considered net benefit, reported positive predictive values (PPV) or directly evaluated clinical impact or utility. Trudell et al.(32) included a brief assessment of cost effectiveness assuming that the model would be used to triage patients for non-stress test (NST) monitoring and that this might reduce stillbirth, although this is not supported by existing evidence.(36) Calibration was rarely reported, and optimism-adjusted calibration measures (i.e. adjusted for overfitting) not considered. Calibration plots, where provided, did suggest overall good calibration,(20,21) as would be expected in development datasets without consideration of overfitting.

DISCUSSION

Main findings

This review identified 69 models predicting stillbirth, none of which were externally validated. There are substantial concerns about risk of bias and applicability precluding the recommendation of any identified model for clinical practice at present. The best apparent performance was reported in models developed for use in later pregnancy incorporating maternal characteristics, placental biomarkers and ultrasound findings.

Several candidate predictors were consistently selected for inclusion in model development and may be important in development of new models. These include ethnicity, maternal BMI, PAPP-A, PlGF and UtAD.

Strengths and limitations

This study provides a broad overview of existing models utilising a comprehensive literature search with a methodologically robust assessment of risk of bias and applicability of included models, in accordance with best practice reporting guidelines.

Direct quantitative comparison of the models included was prevented by the heterogenous predictors and outcomes utilised. A more direct comparison of model performance could be made by external validation of models in an independent dataset. None of the identified models has yet undergone external validation and updating (e.g. recalibration for particular populations). In order to perform independent external validation, the definition of predictors and outcomes and the details of the model algorithm (including intercept and coefficients) are required. Nine models identified are amenable to external validation in suitable datasets.

All of the included models raised concern about either risk of bias or applicability, with a high risk of bias in all but two models. Common concerns related to low EPP, inappropriate modelling of continuous variables or handling of missing data, lack of internal or external validation. Given that few model developers undertook internal validation it was not possible to describe the relationship between model optimism and EPP. The apparent performance of reported models must be considered in the light of lack of adjustment for optimism and incomplete reporting, meaning that most were likely to be overfitted. An overfitted model is unlikely to translate into an effective clinical tool, and may cause harm through inaccurate predicted risks.

Two models were at low risk of bias,(20) but were developed in a low resource setting and less applicable to higher resource settings. These models would be suitable for external validation and clinical appraisal in the setting of intended use.

In all included models, biochemical variables were included as Multiples of the Median (MoMs), commonly used to adjust for laboratory and gestational variance. There have been concerns raised that this adjustment(37) may lead to loss of data and overfitting. Novel model development protocols include these predictors without adjustment.(38)

The predictive accuracy of a model for predicting stillbirth by a fixed gestation over time is affected by the 'competing risk' of live birth. One group undertook time-to-event analysis of the proportional hazards associated with abnormal UtAD pulsatility index,(39) treating live births as censored. This improved discrimination at later gestations, but assumes that censoring was unrelated to prognosis. In fact, women with known risk factors (predictors) for stillbirth are probably systematically delivered earlie, leading to a treatment paradox. This effect is likely present in all routinely collected datasets, and unless accounted for, any prediction model for stillbirth >37 weeks will appear to have limited accuracy. This is critical because this is a key time frame when a prediction model could have significant clinical impact - the most effective intervention to prevent stillbirth remains scheduled birth, usually only pragmatic at term.

Interpretation

The AFFIRM trial tested active management of reduced fetal movement for the prevention of stillbirth but was unable to show a benefit, although intervention increased.(40) This highlights the urgency of accurately identifying women at increased risk in order to minimise iatrogenic harm. Still, developing new models for the prediction of stillbirth is resource intensive, requiring large datasets with high quality information on predictors and interventions. In the seven years since a review identified three models for the prediction of stillbirth,(1) a further 66 have been published. Would the resources required to develop more models be justified when these have yet to be validated, updated and assessed in practice?

Future research should focus on validation and updating of existing first trimester models before subjecting them to clinical evaluation, while development of new second/third trimester models should be a priority. New model development should adhere to reporting guidelines and acknowledge the competing risks inherent pregnancy together with any intervention bias present.

A two-step triage model might be most appropriate - first trimester models with high sensitivity could select a large group for additional monitoring or intervention (e.g. low-dose aspirin). Later pregnancy models could incorporate fetal information (e.g. maternal and fetal Doppler indices and growth) and occurrence of pregnancy complications to give individualised assessment of risk by gestation.

Maternal ethnicity was consistently included as a predictor, but variable definitions lead to concern about generalisability. Ethnicity is consistently associated with both maternal and perinatal mortality, (41,42) but is likely confounded by socio-economic status, structural racism and health literacy. The finding that ethnicity modifies risk of stillbirth is important, and policy makers need this information to underline the importance of increasing population health and equity of access to quality healthcare in reducing stillbirth. Nonetheless, a variable that is inconsistently defined by researchers, applied unpredictably by participants to themselves (43) and increasingly complex with successive generations is arguably inherently unsuitable for precise prediction models.

Stillbirth is a heterogenous outcome related to several pathophysiological pathways. It is implausible that a single test will have high sensitivity for all-cause stillbirth. Prioritisation of sensitivity may lead to clinically useful tests being discarded. This may be best addressed by separate models; logically, the initial target could be placental dysfunction, the largest contributor to global stillbirth. Development of a core outcome set for stillbirth might help to specify outcomes for future models. (44)

Consideration should be given to the timing of the outcome predicted. All included models considered stillbirth as a binary outcome – present or absent at a given gestation. It appeared that the earlier that gestation, the better the performance, but this should be examined in the light of clinical utility. A model predicting stillbirth <32 weeks may have high sensitivity, but is likely to have such a poor PPV that pre-emptive delivery would lead to an unacceptable degree of iatrogenic prematurity. The population incidence affects the PPV of the test, so that even with a high apparent performance the PPV may be as low as 1-3%. The level of risk that justifies intervention is a clinical decision that should be made together with individual women.

Conclusion

This systematic review has identified 69 models incorporating maternal characteristics, biomarkers and ultrasound tests for the prediction of stillbirth. The models identified are at substantial risk of bias

and not yet suitable for use in practice. Future research should focus on the validation of predictive performance (calibration and discrimination) and testing of clinical impact of first trimester models, the development of novel models for use in the third trimester to facilitate individualised mode and timing of birth discussions and development of large, publicly accessible datasets suitable for external validation of existing models. Clinical benefit should also be evaluated, using net benefit and decision curves, and ideally, evaluation of patient outcomes in randomised trials.

Acknowledgements

Jane Sandall is an NIHR Senior Investigator and is also supported by the National Institute for Health Research (NIHR) Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust. The views expressed are those of the author[s] and not necessarily those of the NIHR or the Department of Health and Social Care.

Disclosure of interests

AH reports grants from Tommy's and Action Medical Research, outside the submitted work.
BWM reports grants from the National Health and Medical Research Council (NHMRC) and personal fees from Obseva, Merck, Merck Merck KGaA, Guerbet and iGenomix, outside the submitted work.
GCS reports grants and personal fees from GlaxoSmithKline Research and Development Limited, grants from Sera Prognostics Inc, non-financial support from Illumina Inc, grants, personal fees and non-financial support from Roche Diagnostics Ltd, outside the submitted work. In addition, GCS was a named inventor on a patent application submitted by Cambridge Enterprise for a biomarker test to predict human fetal growth restriction pending.
JS reports support from the NIHR Applied Research Collaboration South London, outside the submitted work.
RT, AK, RR and ST report receipt of a grant from the Stillbirth and Neonatal Death Society (SANDS) during the conduct of this study.
AM, JA, LJ, LM, PvD, MP and KS have no disclosures.
Completed disclosure of interest forms are available to view online as supporting information.

Contribution to authorship

RT planned and carried out the data extraction and analysis and wrote the first draft of the manuscript. AM carried out data extraction and analysis and reviewed and edited the manuscript. JA, KS and RR contributed to study design, interpretation of the results and reviewed and edited the manuscript. ST and AK conceived the project, contributed to study design and reviewed and edited the manuscript. AH, LAM, PvD, GS, JS, BM and BT contributed expertise in stillbirth and prediction model development and reviewed and edited the manuscript. LJ is a patient representative and MP represents parents who have experience of stillbirth via Sands and they consulted on design and reporting of this project.

Details of ethics approval

This was a systematic review of previously published data and as such did not require formal ethics approval.

Funding

The authors are collaborators in the IPPIC (International Prediction of Pregnancy Complications) stillbirth project, funded by Sands (the Stillbirth and Neonatal Death Society).

REFERENCES

1. Kleinrouweler CE, Cheong-See F, Collins G, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol. 2016;214(1):79–90.
2. Blencowe H, Cousens S, Jassir FB, Say L, Chou D, Mathers C, et al. National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. Lancet Glob Heal. 2016 Feb 1;4(2):e98–108.
3. Widdows K, Roberts S, Camacho E, Heazell A. Evaluation of the implementation of the Saving Babies ' Lives Care Bundle in early adopter NHS Trusts in England. Manchester, UK; 2018.
4. Moons KGM, Groot JAH De, Bouwmeester W, Vergouwe Y, Mallett S. Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies : The CHARMS Checklist. PLOS Med. 2014;11(10).
5. Wollf R, Whiting P, Mallett S, Riley R, Westwood M, Kleijnen J, et al. PROBAST: a risk of bias tool for prediction modelling studies. In: Cochrane Colloqium. Vienna; 2015.
6. Riley R, van der Windt D, Croft P. Prognosis Research in Healthcare: Concepts, Methods and Impact. Oxford, UK: Oxford University Press; 2019.
7. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
8. Tavares Da Silva F, Gonik B, McMillan M, Keech C, Dellicour S, Bhange S, et al. Stillbirth: Case definition and guidelines for data collection, analysis, and presentation of maternal

immunization safety data. Vaccine. 2016/07/16. 2016 Dec 1;34(49):6057–68.
9. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017 Jan;356:i6460.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. J Clin Epidemiol. 2009 Oct;62(10):1006–12.
11. Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search Filters for Finding Prognostic and Diagnostic Prediction Studies in Medline to Enhance Systematic Reviews. PLoS One. 2012 Feb 29;7(2):e32844.
12. Ingui BJ, Rogers MA. Searching for clinical prediction rules in MEDLINE. J Am Med Inform Assoc. 2001;8(4):391–7.
13. Steyerberg EW. Clinical Prediction Models. New York, NY: Springer New York; 2009. (Statistics for Biology and Health).
14. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996 Dec 1;49(12):1373–9.
15. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol. 2007 Mar;165(6):710–8.
16. Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, Perneger T V. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. J Clin Epidemiol. 2011 Sep;64(9):993–1000.
17. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. Stat Med. 2018 Oct 22;
18. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016 Nov 24;16(1):163.
19. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med. 2018;(February):1–21.
20. Kayode GA, Grobbee DE, Amoakoh-Coleman M, Adeleke IT, Ansah E, de Groot JAH, et al.

Predicting stillbirth in a low resource setting. BMC Pregnancy Childbirth. 2016;16(1):1–8.
21. Payne BA, Groen H, Ukah UV, Ansermino JM, Bhutta Z, Grobman W, et al. Development and internal validation of a multivariable model to predict perinatal death in pregnancy hypertension. Pregnancy Hypertens. 2015;5:315–21.
22. Akolekar R, Zaragoza E, Poon LCY, Pepes S, Nicolaides KH. Maternal serum placental growth factor at 11 + 0 to 13 + 6 weeks of gestation in the prediction of pre-eclampsia. Ultrasound Obstet Gynecol. 2008 Nov;32(6):732–9.
23. 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;48:631–5.
24. Vellamkondu A, Vasudeva A, Bhat RG, Kamath A, Amin S V., Rai L, et al. Risk Assessment at 11–14-Week Antenatal Visit: A Tertiary Referral Center Experience from South India. J Obstet Gynecol India. 2017;67(6):421–7.
25. Akolekar R, Tokunaka M, Ortega N, Syngelaki A, Nicolaides KH. Prediction of stillbirth from maternal factors, fetal biometry and uterine artery Doppler at 19–24 weeks. Ultrasound Obstet Gynecol. 2016;48(5):624–30.
26. Yerlikaya G, Akolekar R, Mcpherson K, Syngelaki A, Nicolaides KH. Prediction of stillbirth from maternal demographic and pregnancy characteristics. Ultrasound Obstet Gynecol. 2016;48:607–12.
27. Akolekar R, Machuca M, Mendes M, Paschos V, Nicolaides KH. Prediction of stillbirth from placental growth factor at 11–13 weeks. Ultrasound Obstet Gynecol. 2016;48:618–23.
28. Mastrodima S, Akolekar R, Yerlikaya G, Tzelepis T, Nicolaides KH. Prediction of stillbirth from biochemical and biophysical markers at 11 – 13 weeks. Ultrasound Obstet Gynecol. 2016;48:613–7.
29. Åmark H, Westgren M, Persson M. Prediction of stillbirth in women with overweight or obesity—A register-based cohort study. PLoS One. 2018;13(11):1–11.
30. Khalil A, Morales-Rosellõ J, Townsend R, Morlando M, Papageorghiou A, Bhide A, et al. Value of third-trimester cerebroplacental ratio and uterine artery Doppler indices as predictors of stillbirth and perinatal loss. Ultrasound Obstet Gynecol. 2016;47(1).
31. Familiari A, Scala C, Morlando M, Bhide A, Khalil A, Thilaganathan B. Mid-pregnancy fetal growth, uteroplacental Doppler indices and maternal demographic characteristics: role in prediction of stillbirth. Acta Obstet Gynecol Scand. 2016;95(11):1313–8.
32. Trudell AS, Tuuli MG, Colditz GA, Macones GA, Odibo AO. A stillbirth calculator: Development

& internal validation of a clinical prediction model to quantify stillbirth risk. PLoS One. 2017;12(3):1–13.
33. Akolekar R, Bower S, Flack N, Bilardo CM, Nicolaides KH. Prediction of miscarriage and stillbirth at 11 – 13 weeks and the contribution of chorionic villus sampling. Prenat Diagn. 2011;31:38–45.
34. Kingdon C, Roberts D, Turner MA, Storey C, Crossland N, Finlayson KW, et al. Inequalities and stillbirth in the UK: a meta-narrative review. BMJ Open. 2019 Sep 1;9(9):e029672.
35. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Stat Methods Med Res. 2018 Jul 3;096228021878472.
36. Grivell RM, Alfirevic Z, Gyte GML, Devane D. Antenatal cardiotocography for fetal assessment. Cochrane database Syst Rev. 2015 Sep;2015(9):CD007863.
37. Bishop JC, Dunstan FD, Nix BJ, Reynolds TM, Swift A. All MoMs are not equal: some statistical properties associated with reporting results in the form of multiples of the median. Am J Hum Genet. 1993 Feb;52(2):425–30.
38. Mackie FL, Whittle R, Morris RK, Hyett J, Riley RD, Kilby MD. First-trimester ultrasound measurements and maternal serum biomarkers as prognostic factors in monochorionic twins: a cohort study. Diagnostic Progn Res. 2019;3(9):1–9.
39. S Smith GC, H Yu CK, Papageorghiou AT, Maria Cacho A, Nicolaides KH. Maternal Uterine Artery Doppler Flow Velocimetry and the Risk of Stillbirth LEVEL OF EVIDENCE: II. Vol. 109, Obstet Gynecol. 2007.
40. Norman JE, Heazell AEP, Rodriguez A, Weir CJ, Stock SJE, Calderwood CJ, et al. Awareness of fetal movements and care package to reduce fetal mortality (AFFIRM): a stepped wedge, cluster-randomised trial. Lancet. 2018;392(10158):1629–38.
41. Muglu J, Rather H, Arroyo-Manzano D, Bhattacharya S, Balchin I, Khalil A, et al. Risks of stillbirth and neonatal death with advancing gestation at term: A systematic review and meta-analysis of cohort studies of 15 million pregnancies. Smith GC, editor. PLOS Med. 2019 Jul 2;16(7):e1002838.
42. Knight M, Bunch K, Tuffnell D, Shakespeare J, Kotnis R, Kenyon S, et al. Saving Lives, Improving Mothers' Care Lessons learned to inform maternity care from the UK and Ireland Confidential Enquiries into Maternal Deaths and Morbidity 2015-17. Oxford, UK; 2019.
43. Lockie E, Mccarthy EA, Hui L, Churilov L. Feasibility of using self-reported ethnicity in

pregnancy according to the gestation-related optimal weight classification : a cross-sectional study. BJOG. 2018;125:704–9.

44. Duffy JMN, Ziebland S, von Dadelszen P, McManus RJ. Tackling poorly selected, collected, and reported outcomes in obstetrics and gynecology research. Am J Obstet Gynecol. 2019 Jan 1;220(1):71.e1-71.e4.

Figures and Tables

Online Supporting Material

Table 1. Predictors included in the final models provided in each paper

Study	Maternal characteristics					Maternal history													Biochemical tests			Ultrasound tests			
	Age	Weight	Height	Body mass index	Ethnicity	Maternal co-morbidities	Mode of Conception	Smoking	Alcohol	Hypertension	Diabetes	APS/SLE	Maternal Symptoms	Previous fetal loss	Parity	Place of Residence	Education/Occupation	Bleeding	Fetal Presentation	Placental growth factor	PAPP-A	Proteinuria	Fetal Biometry	UtAPI	Ductus venosus	Cerbroplacental ratio
Akolekar 2011	•	•	•		•			•		•					•						•				•	
Akolekar 2016 (1)		•			•		•	•		•	•	•		•									•	•		
Akolekar 2016 (2)		•			•		•	•		•	•	•		•						•				•	•	
Aupont 2016		•			•		•	•		•	•	•		•						•			•	•		
Amark 2018	•			•	•			•							•						•					
Familiari 2016	•			•	•																		•	•		
Kayode 2016						•								•	•	•	•	•	•				•			
Khalil 2016																							•	•		•
Mastrodima 2016		•			•		•	•		•	•	•		•							•			•	•	
Payne 2015	•												•									•				
Smith 2007				•	•																			•		
Trudell 2017	•			•	•			•		•	•				•											
Vellamkondu 2017	•													•							•					
Yerlikaya 2016		•			•		•	•		•	•	•		•												

Figure 1. PRISMA flow diagram

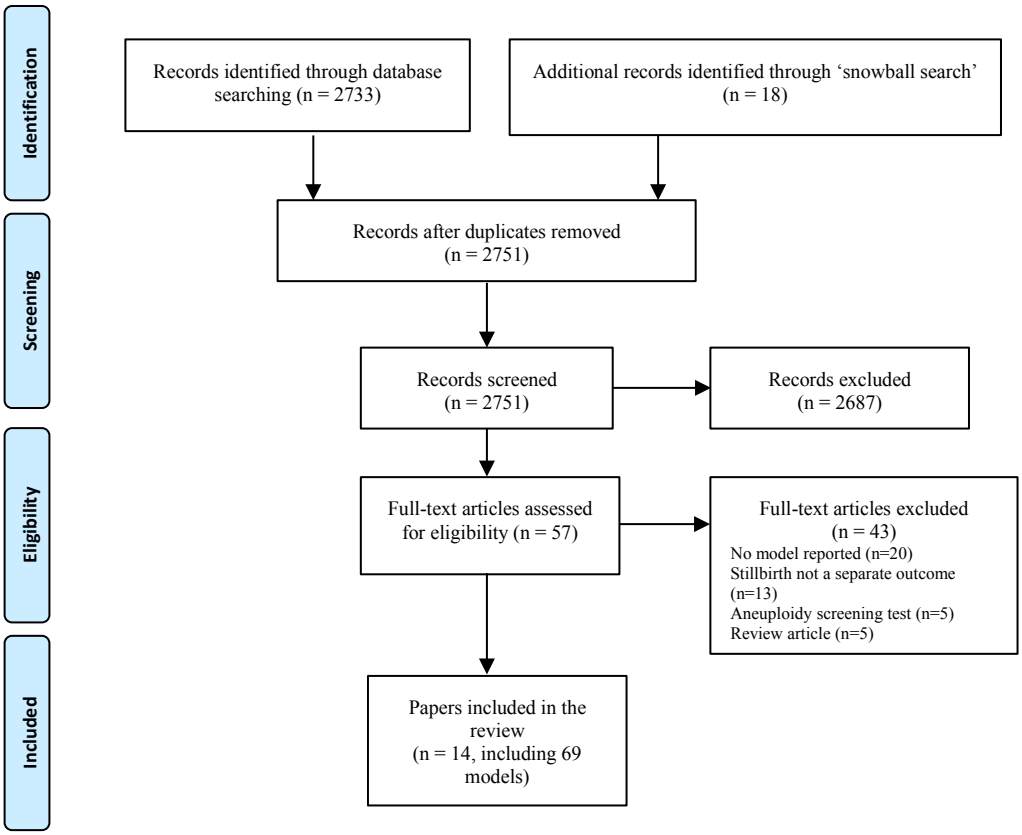

Figure 2a. PROBAST risk of bias

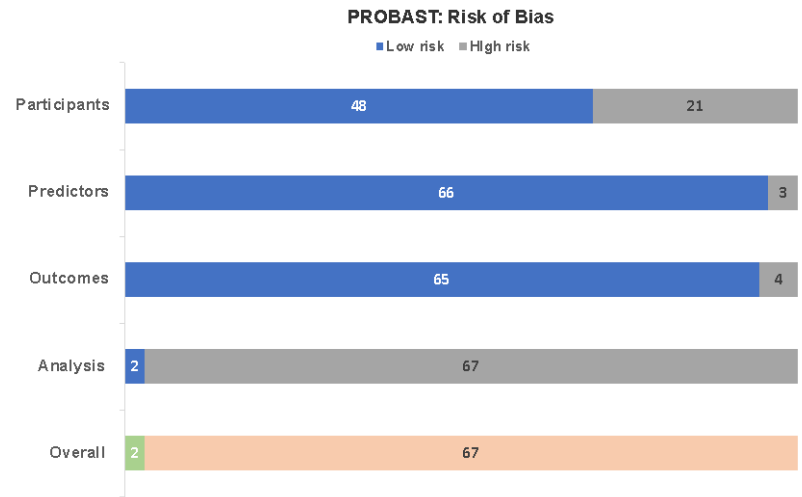

Figure 2b. PROBAST Applicability

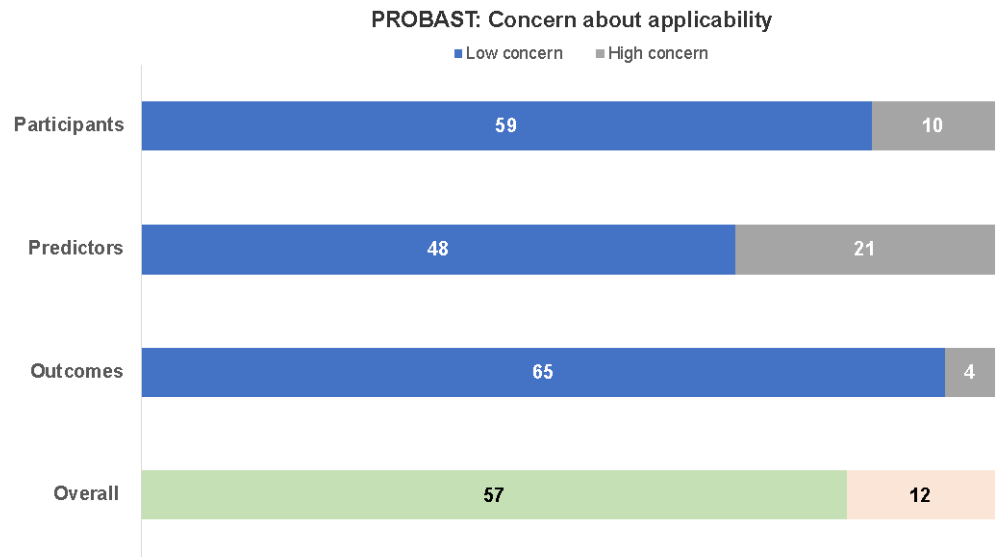

Figure 3a. Model performance for "all stillbirth" outcome

The figure shows the apparent area under the curve (AUC) with their 95% confidence intervals reported for models predicting "all stillbirth" or "all antepartum stillbirth"

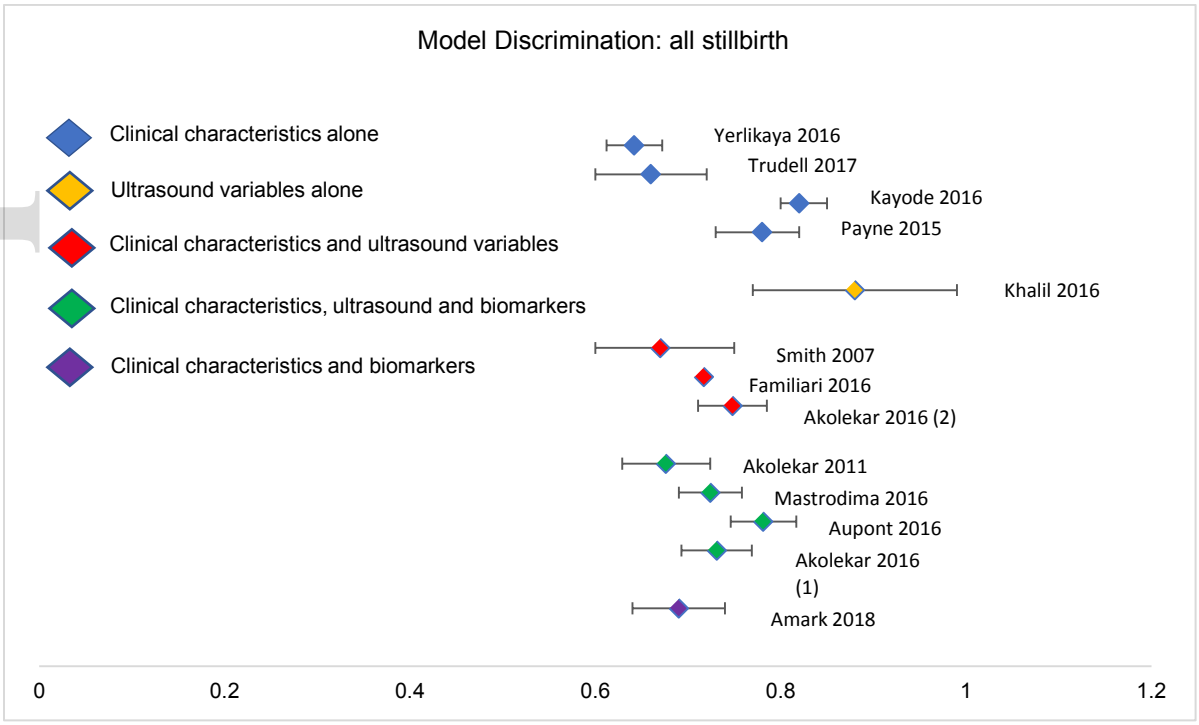

Figure 3b. Model performance for stillbirth at term

The figure shows the area under the curved (AUC) with their 95% confidence intervals reported for models predicting stillbirth >37 weeks