AOGS
Acta Obstetricia et Gynecologica Scandinavica

AOGS SYSTEMATIC REVIEW

Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation

LINDA J.E. MEERTENS[1,*], PIM VAN MONTFORT[1,*], HUBERTINA C.J. SCHEEPERS[2], SANDER M.J. VAN KUIJK[3], ROBERT AARDENBURG[4], JOSJE LANGENVELD[4], IVO M.A. VAN DOOREN[5], IRIS M. ZWAAN[6], MARC E.A. SPAANDERMAN[2] & LUC J.M. SMITS[1]

[1]Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, [2]Department of Obstetrics and Gynecology, School for Oncology and Developmental Biology (GROW), Maastricht University Medical Center, Maastricht, [3]Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA), Care and Public Health Research Institute (CAPHRI), Maastricht University Medical Center, Maastricht, [4]Department of Obstetrics and Gynecology, Zuyderland Medical Center, Heerlen, [5]Department of Obstetrics and Gynecology, Sint Jans Gasthuis Weert, Weert, and [6]Department of Obstetrics and Gynecology, Laurentius Hospital, Roermond, The Netherlands

Key words
Spontaneous preterm birth, spontaneous preterm delivery, prediction, risk assessment, external validation, decision curve analysis, systematic review

Correspondence
Pim van Montfort, Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, PO Box 616, 6200 MD, Maastricht, The Netherlands.
E-mail: pim.vanmontfort@maastrichtuniversity.nl

Conflict of interest
The authors have stated explicitly that there are no conflicts of interest in connection with this article.
*Contributed equally.

Please cite this article as: Meertens LJE, van Montfort P, Scheepers HCJ, van Kuijk SMJ, Aardenburg R, Langenveld J et al. Prediction models for the risk of spontaneous preterm birth based on maternal characteristics: a systematic review and independent external validation. Acta Obstet Gynecol Scand. 2018; 97:907–920.

Received: 24 January 2018
Accepted: 10 April 2018

DOI: 10.1111/aogs.13358

Abstract

Introduction. Prediction models may contribute to personalized risk-based management of women at high risk of spontaneous preterm delivery. Although prediction models are published frequently, often with promising results, external validation generally is lacking. We performed a systematic review of prediction models for the risk of spontaneous preterm birth based on routine clinical parameters. Additionally, we externally validated and evaluated the clinical potential of the models. *Material and methods.* Prediction models based on routinely collected maternal parameters obtainable during first 16 weeks of gestation were eligible for selection. Risk of bias was assessed according to the CHARMS guidelines. We validated the selected models in a Dutch multicenter prospective cohort study comprising 2614 unselected pregnant women. Information on predictors was obtained by a web-based questionnaire. Predictive performance of the models was quantified by the area under the receiver operating characteristic curve (AUC) and calibration plots for the outcomes spontaneous preterm birth <37 weeks and <34 weeks of gestation. Clinical value was evaluated by means of decision curve analysis and calculating classification accuracy for different risk thresholds. *Results.* Four studies describing five prediction models fulfilled the eligibility criteria. Risk of bias assessment revealed a moderate to high risk of bias in three studies. The AUC of the models ranged from 0.54 to 0.67 and from 0.56 to 0.70 for the outcomes spontaneous preterm birth <37 weeks and <34 weeks of gestation, respectively. A subanalysis showed that the models discriminated poorly (AUC 0.51–0.56) for nulliparous women. Although we recalibrated the models, two models retained evidence of overfitting. The decision curve analysis showed low clinical benefit for the best performing models. *Conclusions.* This review revealed several reporting and methodological shortcomings of published prediction models for spontaneous preterm birth. Our external validation study indicated that none of the models had the ability to predict spontaneous preterm birth adequately in our population. Further improvement of prediction models, using recent knowledge about both model development and potential risk factors, is necessary to provide an added value in personalized risk assessment of spontaneous preterm birth.

Abbreviations: AUC, area under the receiver operating characteristic curve; BMI, body mass index; CHARMS, CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies; CI, confidence interval; PTB, preterm birth; sPTB, spontaneous preterm birth; TRIPOD, transparent reporting of a multivariable prediction model for individual prognosis or diagnosis.

Introduction

Preterm birth (PTB), usually defined as birth before 37 weeks of gestation, occurs in 5–10% of singleton pregnancies in Europe (1). The majority of preterm deliveries, approximately 70%, start spontaneously (sPTB) (2). As both perinatal mortality and morbidity are inversely related to gestational age, health benefits may be achieved by increased monitoring and preventive interventions, resulting in a prolongation of pregnancy (3,4).

Progesterone treatment has been reported to reduce the risk of sPTB before 34 weeks of gestation in women at high risk (5,6). Cervical cerclage or application of a pessary may also protect against sPTB (7–9). Evidence on which of the three interventions is most effective is limited (7–9).

Women with a history of sPTB, cervical surgery or a mid-pregnancy short cervix are considered to be at high risk (10). Without routine cervical length screening, the majority of nulliparous women are regarded as low risk and thus do not receive any preventive treatment. However, universal cervical length screening in women without a history of sPTB results in relatively high numbers needed to screen (1147 in low-risk nulliparous women) (11,12). Universal cervical length screening is not performed in Dutch obstetric care. Besides a history of sPTB, other risk factors have been associated with PTB, including socioeconomic status, psychological characteristics, family history, height, weight, and smoking (13). Early risk assessment may be useful to identify women at risk who may benefit from effective follow-up management strategies.

In the past, several risk assessment tools for sPTB based on a list of single risk factors were developed showing low accuracy rates (14). In the last decade, a number of promising prediction models based on multivariable regression analysis for the risk of sPTB have been published (15). Prediction models may be more accurate in identifying women at high risk, as regression allows for a more fine-tuned estimation of the weight of multiple risk factors and possible inter-relations (16). A review of all existing models assessing their methodological quality is lacking. Moreover, most models have not been externally validated, an essential step before implementation in clinical practice (17). In this article, we performed a systematic review of all existing models predicting sPTB based on routine clinical parameters obtained in first 16 weeks of pregnancy. We externally validated and compared the selected models in a Dutch multicenter prospective cohort of pregnant women.

Material and methods

Data sources

This systematic review is reported in accordance with the recently published guidelines for systematic reviews and meta-analyses of prediction model performance (18). We systematically searched PubMed and EMBASE up to 26 June 2017. Keywords for prediction studies were combined with synonyms for the outcome sPTB appearing in the title, abstract or MeSH terms. Reference lists of included studies and related articles (i.e. reviews) were manually checked to identify additional eligible articles. The detailed search strategy is provided as Supporting Information Appendix S1.

Eligibility criteria

We aimed to identify all published prediction models for the risk of sPTB that are applicable in the first 16 weeks of pregnancy and are based on non-invasive predictors (Appendix S1). Studies were eligible if they met the following criteria: (i) the article presented a newly developed

Key Message

Prediction models may contribute to personalized risk-based management of women at high risk of spontaneous preterm delivery. This systematic review identified promising non-invasive prediction models. However, external validation indicated that no model could adequately predict spontaneous preterm birth.

prediction model, or a validation or update of a previously developed model in pregnant women, (ii) the outcome of the model was the risk of sPTB, (iii) the model contained more than one predictor, (iv) predictors were available in Dutch obstetric practice (maternal characteristics, anthropometric measures or blood pressure measurements), (v) predictor values were obtainable during first 16 weeks of pregnancy, (vi) these predictor values were based on regression coefficients. Authors of the original articles were contacted if the model algorithm or definitions of predictors were not available. Studies were excluded when the regression coefficients could not be obtained, the paper was written in a language other than English, German, French or Dutch, or if it was a non-original study (i.e. review). Two researchers (L.M., P.v.M.) screened the retrieved titles and abstracts and assessed the eligibility of the full-text papers independently. Discrepancies were resolved by discussion. A third reviewer (L.S.) was available in case no consensus was reached.

The risk of bias of the included studies was assessed using the CHecklist for critical Appraisal and data extraction for systematic Reviews of prediction Modelling Studies (CHARMS) (19). The following data were extracted for each included study: source of data, participants, outcome(s) to be predicted, candidate predictors, sample size, handling of missing data, model development, model performance, model evaluation, model presentation and model interpretation. The risk of bias was critically assessed for eight risk domains: source of data, participant selection, predictor assessment, outcome assessment, sample size, attrition, analysis, and presentation of the model. Risk of bias was rated as low if bias was unlikely, moderate if there were no fatal shortcomings, and high if essential errors were made. Previously published risk of bias criteria were used and slightly adapted (20). Data extraction and critical appraisal were performed independently by two reviewers (L.M., P.v.M.). Discrepancies were resolved by discussion and a third reviewer (L.S.) was available in case no consensus was reached.

Data collection and analysis

The included prediction models were externally validated in the Expect Study I (21). The main purpose of the Expect Study I was to validate published prediction models for several obstetric complications in an independent population. A multicenter prospective cohort study was performed in 36 midwifery practices (primary care) and six hospitals (secondary and tertiary care) in the south-eastern part of The Netherlands between 1 July 2013 and 1 January 2015. The patients were followed up until 31 December 2015. All pregnant women up to 16 weeks of gestation and aged 18 years or older were eligible. Eligible pregnant women were asked to complete two web-based questionnaires (a paper version was available upon request), one before 16 weeks of gestation and the other 6 weeks after the estimated due date. The online questionnaires were accessible via the study website using a unique login code provided with the study information. Automatic reminders were sent in the case of incompleteness or nonresponse. Medical records and discharge letters were requested from caregivers. Pregnancies ending in a miscarriage or termination before 24 weeks of gestation, and women lost to follow up, were excluded. For this study, we also excluded multiple pregnancies and cases of iatrogenic preterm onset of parturition.

Predictors in the included prediction models were assessed by the pregnancy questionnaire completed before 16 weeks of gestation. We used the same definitions as defined in the original articles (Supporting Information Appendix S2).

The primary outcome sPTB was defined as a delivery before 37 weeks of gestation with spontaneous onset of parturition (primary contractions or preterm premature rupture of membranes). Secondly, we defined early sPTB as a spontaneously delivery before 34 weeks of gestation. The outcome was obtained from a combination of the medical record and postpartum questionnaire. Cause of labor onset (i.e. spontaneous or not) was available in both data sources. Duration of pregnancy was also available in both data sources and was moreover calculated based on estimated due date and date of birth. Discrepancies between the two variables and data sources were checked. In the absence of the postpartum questionnaire (n = 421 sPTB <37 weeks and n = 424 sPTB <34 weeks), the medical record was used as reference standard and vice versa (n = 16 for both sPTB <37 weeks and sPTB <34 weeks).

A sample size of 2500 women was expected to provide a minimum of 100 cases and 100 non-cases, assuming a 4.5% incidence rate of sPTB <37 weeks of gestation (22).

We imputed missing data for predictors using stochastic regression imputation with predictive mean matching as the imputation model (23). Characteristics of the validation cohort were described as an absolute value (percentage) for categorical variables and as mean ± standard deviation (SD) for continuous variables. We evaluated the relatedness of development samples and validation cohort by comparing the distribution of population characteristics.

The original formulas were used to calculate individual predicted probabilities for each model (Appendix S2). We assessed the predictive performance of each model by means of discrimination and calibration for the outcomes sPTB <37 and <34 weeks of gestation, as described in the framework reported by Steyerberg et al. (16). Discrimination indicates the ability of the model to distinguish between women who will have a sPTB and those who will not. For each model, we computed the area under the receiver operating characteristic curve (AUC) with a 95% confidence interval (CI). A subgroup analysis was performed among nulliparous women, as a history of sPTB is a strong risk factor for recurrent sPTB. Calibration refers to the agreement between the actual outcomes and predicted probabilities by the model. We constructed calibration plots in which women were divided into 10 groups with similar predicted risks, and calculated calibration-in-the-large and the slope. Calibration-in-the-large (intercept), which compares the mean predicted probabilities with mean observed risk, indicates the extent to which predictions are systematically too low or too high. The slope refers to the average strength of predictor effects. Perfect predictions have an intercept of zero and a slope of one (17). The prediction models were recalibrated by adjusting the intercept and slope using the linear predictor as the only covariate. Discriminative performance (AUC) of the models is not affected, as this recalibration method does not change the ranking of the predicted probabilities (24). A discriminative performance below 0.70 is generally considered moderate (16).

Lastly, we performed decision curve analysis to evaluate the potential clinical utility of the models. Decision curve analysis assesses the net benefit (proportion of true-positives and false-positives) of the prediction models over a range of risk thresholds compared with considering all and no women to be at high risk for sPTB (25). Sensitivity, specificity, and positive and negative predictive values at certain risk thresholds were calculated for the model with the highest overall net benefit.

Statistical analyses were performed with R version 3.4.1, packages rms, pROC, and DecisionCurve.

Results

General characteristics of the studies

The search identified 2018 unique articles. After title and abstract screening, full text assessment was performed for 47 articles. Four articles fulfilled the eligibility criteria (26–29). Reference cross-checking provided no additional articles. An overview of the systematic study selection is shown in Appendix S1.

The four included studies were all development studies describing five models predicting the risk for sPTB based on maternal characteristics. The studies were conducted in four different countries and published between 2011 and 2014. Two studies used a prospective cohort design and the other two were based on registry data. The number of predictors in the published prediction models varied between 2 and 16. Common predictors were body mass index (BMI), smoking and previous PTB. The prevalence of sPTB, defined as sPTB <34 weeks of gestation by two studies and <37 weeks of gestation by the other two studies, ranged from 0.9 to 1.1% for sPTB <34 weeks of gestation and from 3.7 to 5.7% for sPTB <37 weeks of gestation. Discriminative performance (AUC) varied from 0.62 to 0.70. Only one study performed internal validation by bootstrapping and the study of Sananes et al. (26) performed an external validation of which the results were not reported. The key characteristics of the included studies are shown in Table 1.

A summary of potential bias per domain is shown in Figure 1. Two studies used registry data for model development, which may be less effective for research purposes due to the likelihood of missing data on promising predictors. Moreover, the outcome was extracted at the same time as the predictors, which may lead to bias. Nevertheless, sPTB is an objective outcome, so assessment may be less biased. The domain participants was rated as liable to a moderate to high risk of bias due to selective reporting of patient characteristics. Parra-Cordero et al. (28) used criteria which are not available at the intended moment of prediction. Besides, women may be treated for spontaneous onset of PTB. Only Alleman et al. (27) explicitly reports exclusion of women undergoing cerclage or tocolysis from their study population. Parra-Cordero et al. (28) merely excluded women with a history of cerclage. Sample size was scored at moderate risk for the model of Parra-Cordero et al. (28) because the overall number of cases was low (n = 31), which probably led to the inclusion of only two predictors. The domains attrition and analysis had the highest risk of bias for all included models. All studies either had incomplete data (lost-to-follow-up or missing predictor values), or did not report any information about missing data [Parra-Cordero et al. (28)]. The other three studies were scored as moderate risk because they had a substantial amount of missing data and performed a complete case analysis. Methods of analysis were not reported in enough detail by Parra-Cordero et al. (28). All studies selected predictors based on statistical significance and only one study performed shrinkage of the regression coefficients. For the models of two studies, only odds ratios were available. As the intercept was unavailable, no initial calibration plots could be drawn. Alleman et al. (27) reported their final model including serum markers. The algorithm consisting only maternal characteristics was provided after contacting the authors. Overall, the study of Beta et al. (29) showed the lowest risk of bias. A detailed description of the data extraction and risk of bias assessment according to the CHARMS checklist is provided in Supporting Information Appendix S3.

Table 1. Characteristics included prediction models for spontaneous preterm birth.

Study, Author (year)	Study design	Population	Time of assessment	No. cases/total (%)	Definition sPTB	Predictors	Prediction model
Parra-Cordero et al. (2014)	Prospective cohort (*n* = 3480)	Singleton pregnancies Exclusion: iatrogenic delivery <34 weeks of gestation, early-onset preeclampsia, early-onset SGA, spontaneous miscarriage, intrauterine fetal death, fetal abnormalities, placental abruption, cerclage, and history of cervical surgery	11^{+0} to 13^{+6} weeks of gestation	31/3310 (0.9)	sPTB <34 weeks of gestation	Prior preterm delivery, smoking	Odds ratios reported
Sananes et al. (2013)*	Registry data 2000–2011 (*n* = 33 761)	Singleton pregnancies Exclusion: fetal deaths, medical terminations, iatrogenic delivery <37 weeks of gestation, and delivery <24 weeks of gestation	<14 weeks of gestation	NR/17,341 (NR)	sPTB <37 weeks of gestation	Age, BMI, prior late miscarriage, prior preterm delivery, prior term delivery, smoking	Odds ratios reported. Full algorithm received by email
Alleman et al. (2013)*	Registry data 2009–2010 (*n* = 12 057)	Singleton pregnancies Exclusion: congenital anomaly, birthweight >3 SD from mean, serious infection, cerclage, tocolysis, and delivery <20 weeks of gestation	First trimester (precise period NR)	153/2699 (5.7)	sPTB <37 weeks of gestation	BMI, diabetes mellitus, education, prior preterm delivery, prior live birth	Algorithm and odds ratios not reported. Full algorithm received by email
Beta et al. (2011)	Prospective cohort 2006–2009 (*n* = 36 743)	Singleton pregnancies Exclusion: major fetal abnormalities, termination, miscarriage or fetal death before 24 weeks of gestation, and iatrogenic delivery <34 weeks of gestation	11^{+0} to 13^{+6} weeks of gestation	353/33 370 (1.0)	sPTB <34 weeks of gestation	Age, ethnicity, height, method of conception, nulliparous fetal loss, nulliparous late miscarriage, prior preterm birth, prior iatrogenic preterm delivery, prior term delivery, smoking Model 2 with obstetric history subdivided according to number of previous preterm deliveries	Odds ratios reported.

AUC, area under the receiver operating characteristic curve; BMI, body mass index; CI, confidence interval; NR, not reported; sPTB, spontaneous preterm birth.
*Sananes et al. (26) externally validated their model, but results of this validation are not reported. Alleman et al. (27) performed an internal validation step by 1000-fold bootstrapping, but did not report the results.

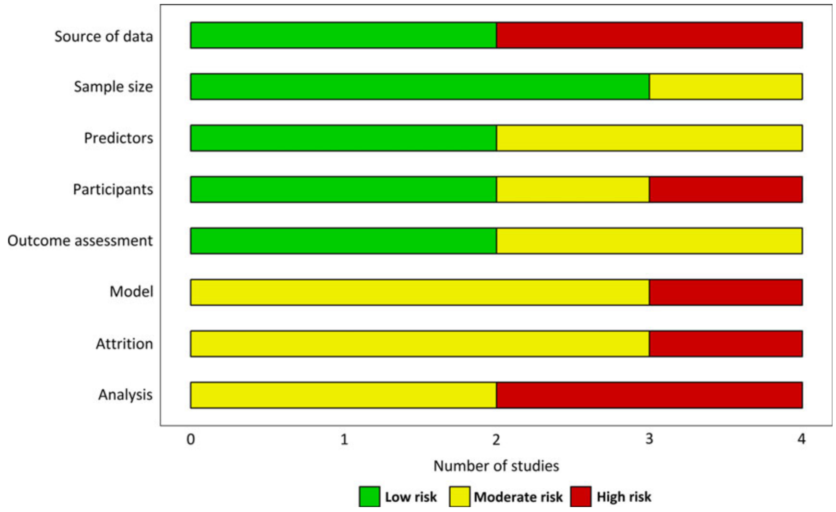

Figure 1. Risk of bias assessment of the four included studies according to CHARMS checklist (19). [Color figure can be viewed at wileyonlinelibrary.com]

Synthesis of the results

The validation cohort consisted of 2540 women of which 118 (4.6%) had an sPTB <37 weeks of gestation (Figure 2). Patient characteristics are shown in Table 2. There were ≤1.2% missing values per predictor and the cohort was generally similar after imputation of incomplete predictor variables. Supporting Information Appendix S4 provides an overview of complete cases and the imputed validation cohort. The study population for the outcome sPTB <34 weeks of gestation comprised 2576 women, since fewer women were excluded because of an iatrogenic preterm onset of labor, of which 34 women (1.3%) delivered spontaneously before 34 weeks of gestation.

The distribution of predictors and predictor effects in the original cohorts and our validation cohort are available in Supporting Information Appendix S5. In contrast to the original cohorts, women in our validation cohort were nearly all of Caucasian origin. Almost all population characteristics of Sananes et al. (26) differed considerably compared with the validation cohort. Women in the cohort of Alleman et al. (27) had a higher BMI and higher prevalence of pre-existing diabetes mellitus. The populations of Parra-Cordero et al. (28) and Beta et al. (29) were more comparable, but Parra-Cordero et al. (28) had a higher prevalence of smoking during pregnancy and women in the cohort of Beta et al. (29) were shorter and had a higher prevalence of previous fetal loss. The prevalence of sPTB <37 weeks of gestation was higher in Alleman et al. (27) (5.7%) and lower in the overall population of Sananes et al. (26) (3.7%) compared with the validation cohort (4.6%). The outcome sPTB <34 weeks of gestation was comparable with our prevalence.

The discriminative performance of the included models is shown in Table 3. For the primary outcome sPTB <37 weeks of gestation, the AUC ranged from 0.54 to 0.67. The AUC of the model of Alleman et al. (27) decreased considerably from 0.70 to 0.57 [95% CI 0.52,0.62]. The model of Sananes et al. (26) had a slightly higher discrimination compared with the original cohort. All models performed better for the outcome sPTB <34 weeks of gestation. Model 2 of Beta et al. (29) yielded the highest discriminative performance [AUC 0.70, 95% CI 0.61–0.78]. Wide confidence intervals were observed due to the low number of cases for sPTB <34 weeks of gestation. The subgroup analysis among nulliparous women showed a drastic decrease towards almost no discriminative performance for all models. The receiver operating characteristic (ROC) curves in the overall cohort are presented in Figure 3.

Calibration plots of the two models that provided a complete algorithm are provided in Figure 4. The model of Alleman et al. (27) underestimated the risk of sPTB and was overfitted (slope <1). Besides the difference in baseline risk, Sananes et al. (26) fitted well with our population (slope = 1). Recalibration showed closer fitting to the ideal calibration line (Supporting Information Appendix S6). The models of Alleman et al. (27) and Beta et al. (29) retained some overfitting.

The decision curve analysis of the two best performing models is presented in Figure 5. The models had a positive net benefit compared with classifying all or no women as high risk over a small range of probability thresholds (2.5–10%). However, net benefit remained low throughout this range. This low clinical usefulness is also shown in Table 4. Choosing a high sensitivity leads to a

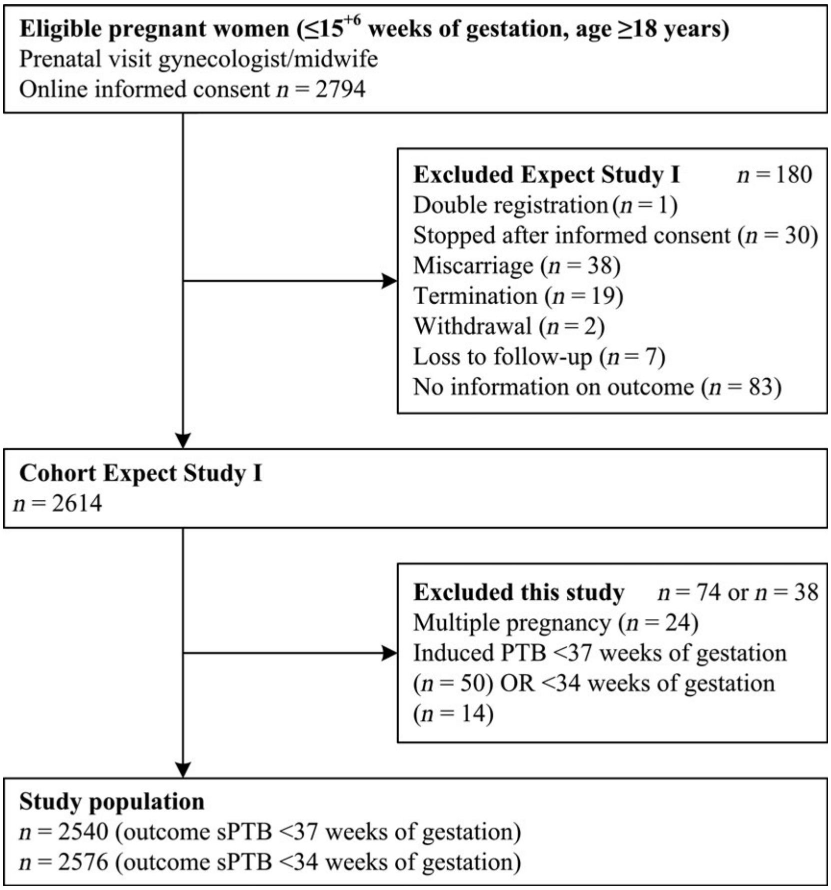
Eligible pregnant women (≤15+6 weeks of gestation, age ≥18 years)
Prenatal visit gynecologist/midwife
Online informed consent n = 2794
Excluded Expect Study I n = 180
Double registration (n = 1)
Stopped after informed consent (n = 30)
Miscarriage (n = 38)
Termination (n = 19)
Withdrawal (n = 2)
Loss to follow-up (n = 7)
No information on outcome (n = 83)
Cohort Expect Study I
n = 2614
Excluded this study n = 74 or n = 38
Multiple pregnancy (n = 24)
Induced PTB <37 weeks of gestation (n = 50) OR <34 weeks of gestation (n = 14)
Study population
n = 2540 (outcome sPTB <37 weeks of gestation)
n = 2576 (outcome sPTB <34 weeks of gestation)

Figure 2. Flowchart validation cohort spontaneous preterm birth (sPTB).

large proportion of women who will be indicated unnecessarily as having a high risk of sPTB <37 weeks of gestation. Conversely, a higher specificity leads to a minimal amount of true-positives. The model was especially insufficient among nulliparous women. The moderate performance is predominantly determined by a history of sPTB or term delivery.

Ethical approval

The Medical Ethical Committee of the Maastricht University Medical Center evaluated the study protocol and declared that no ethical approval was necessary (MEC 13-4-053). All participating women gave informed consent via the Internet. The study was registered at The Netherlands Trial Registry on 21 August 2013 (NTR4143, www.trialregister.nl).

Discussion

In this systematic review we provide an overview of the currently available prediction models of sPTB based on routine clinical parameters. We identified four articles describing five models fulfilling the eligibility criteria. Assessment of methodological quality revealed several shortcomings in the reporting of models. Furthermore, there is a moderate to high risk of bias in the development of the models according to the CHARMS criteria. External validation resulted in a decreased discriminative ability for all models. Model 2 of Beta et al. (29) had the highest AUC (sPTB <37 weeks: 0.67, and sPTB <34 weeks: 0.70) after validation. This model was based on age, ethnicity, height, method of conception, nulliparous fetal loss, nulliparous late miscarriage, prior PTB (subcategories), prior iatrogenic PTB, prior term delivery and smoking. The model of Sananes et al. (26) showed the best calibration (slope of one) for sPTB <37 weeks of gestation.

Our systematic review identified a moderate reporting quality of most studies according to the CHARMS criteria. Reporting shortcomings were also noted in a general systematic review of obstetric prediction models (15). The recently published transparent reporting of a multivariable prediction model for individual prognosis or

Table 2. Baseline characteristics of the validation cohort (Expect Study I).

		Observed validation cohort (Expect Study I)[a]		
Characteristics	Missing values, *n* (%)	Overall (*n* = 2540)	sPTB <37 weeks (*n* = 118)	No sPTB ≥37 weeks (*n* = 2422)
Age, years	0 (0.0)	30.2 (3.9)	30.1 (3.8)	30.2 (3.9)
Ethnicity				
Caucasian	0 (0.0)	2462 (96.9)	115 (97.5)	2347 (96.9)
Afro-Caribbean		3 (0.1)	1 (0.8)	2 (0.1)
South Asian		4 (0.2)	0 (0.0)	4 (0.2)
East Asian		4 (0.2)	1 (0.8)	3 (0.1)
Other Asian		11 (0.4)	1 (0.8)	10 (0.4)
Hispanic		11 (0.4)	0 (0.0)	11 (0.5)
Mixed		45 (1.8)	0 (0.0)	45 (1.9)
Tertiary level of education	3 (0.1)	1380 (54.3)	69 (58.5)	1311 (54.1)
Height, cm	3 (0.1)	168.8 (6.4)	167.3 (6.6)	168.9 (6.4)
Weight, kg	5 (0.2)	68.9 (13.0)	65.6 (11.5)	69.0 (13.0)
Body mass index, kg/m^2	5 (0.2)	24.1 (4.3)	23.4 (3.8)	24.2 (4.3)
Smoking during pregnancy	1 (0.0)	149 (5.9)	8 (6.8)	141 (5.8)
Diabetes mellitus	0 (0.0)	10 (0.4)	1 (0.8)	9 (0.4)
Type 1		8 (0.3)	1 (0.8)	7 (0.3)
Type 2		1 (0.0)	0 (0.0)	1 (0.0)
Other		1 (0.0)	0 (0.0)	1 (0.0)
History of chronic hypertension	0 (0.0)	24 (0.9)	0 (0.0)	24 (1.0)
Parity				
Nulliparous	0 (0.0)	1284 (50.6)	77 (65.3)	1207 (49.8)
Primiparous		1003 (39.5)	35 (29.7)	968 (40.0)
Multiparous		253 (9.9)	6 (5.0)	247 (10.2)
Conception				
Spontaneous	0 (0.0)	2375 (93.5)	114 (96.6)	2261 (93.4)
Ovulation induction		88 (3.5)	3 (2.5)	85 (3.5)
IVF/ICSI		77 (3.0)	1 (0.8)	76 (3.1)
History of fetal loss <16 weeks of gestation	0 (0.0)	702 (27.6)	24 (20.3)	678 (28.0)
History of recurrent miscarriages (≥3)	0 (0.0)	49 (1.9)	1 (0.8)	48 (2.0)
Vaginal bleeding (≥2 days)	0 (0.0)	277 (10.9)	27 (20.3)	250 (10.3)
History of sPTB	30 (1.2)	76 (3.0)	16 (13.6)	60 (2.5)
16–23 weeks of gestation		4 (0.2)	1 (0.8)	3 (0.1)
24–27 weeks of gestation		7 (0.3)	1 (0.8)	6 (0.2)
28–30 weeks of gestation		2 (0.1)	2 (1.7)	0 (0.0)
31–33 weeks of gestation		13 (0.5)	3 (2.5)	10 (0.4)
34–36 weeks of gestation		52 (2.0)	9 (7.6)	43 (1.8)
History of iatrogenic preterm delivery ≥24 weeks of gestation	29 (1.1)	44 (1.7)	0 (0.0)	44 (1.8)
History of term delivery	29 (1.1)	1130 (44.5)	29 (24.6)	1101 (45.5)
History of live birth	18 (0.7)	1221 (48.1)	40 (33.9)	1181 (48.8)

ICSI, intracytoplasmic sperm injection; IVF, in vitro fertilization; sPTB, spontaneous preterm birth.

[a] Original data (not imputed) presented as mean (SD) or absolute number (%).

diagnosis (TRIPOD) statement, may lead to improvements in the reporting quality of future studies (30).

Risk of bias assessment revealed a moderate to high risk of bias in three of four studies. The main sources of bias were in the domains of analysis, attrition and modeling. All studies selected predictors on the basis of statistical significance, which leads to a model that fits the data too closely (24,31). Next, continuous variables were often dichotomized, for example age and BMI in two of our selected models, leading to loss of information (32).

Table 3. Discrimination of selected prediction models for spontaneous preterm birth.

Study, Author (year)	Discrimination C-Statistic [95% CI] Original publication	Discrimination C-Statistic [95% CI] Validation cohort sPTB <37 weeks (*n* = 2540)	Discrimination C-Statistic [95% CI] Validation cohort sPTB <34 weeks (*n* = 2576)	Discrimination C-Statistic [95% CI] Validation cohort, nulliparous sPTB <37 weeks (*n* = 1284)	Discrimination C-Statistic [95% CI] Validation cohort, nulliparous sPTB <34 weeks (*n* = 1305)
Parra-Cordero et al. (2014)	NR	0.54 [0.50,0.57]	0.56 [0.49,0.63]	0.52 [0.50,0.54]	0.51 [0.46,0.55]
Sananes et al. (2013)	0.618 [0.595,0.641]	0.64 [0.60,0.68]	0.68 [0.59,0.76]	0.53 [0.48,0.57]	0.53 [0.43,0.63]
Alleman et al. (2013)	0.703 [NR]	0.57 [0.52,0.62]	0.61 [0.51,0.71]	0.55 [0.49,0.60]	0.51 [0.39,0.63]
Beta et al. (2011)	Model 1: 0.668 [0.639,0.698] Model 2: NR	0.65 [0.60,0.70] 0.67 [0.62,0.72]	0.68 [0.59,0.77] 0.70 [0.61,0.78]	0.51 [0.45,0.57] 0.54 [0.48,0.60]	0.52 [0.39,0.65] 0.56 [0.44,0.68]

CI, confidence interval; NR, not reported; sPTB, spontaneous preterm birth.

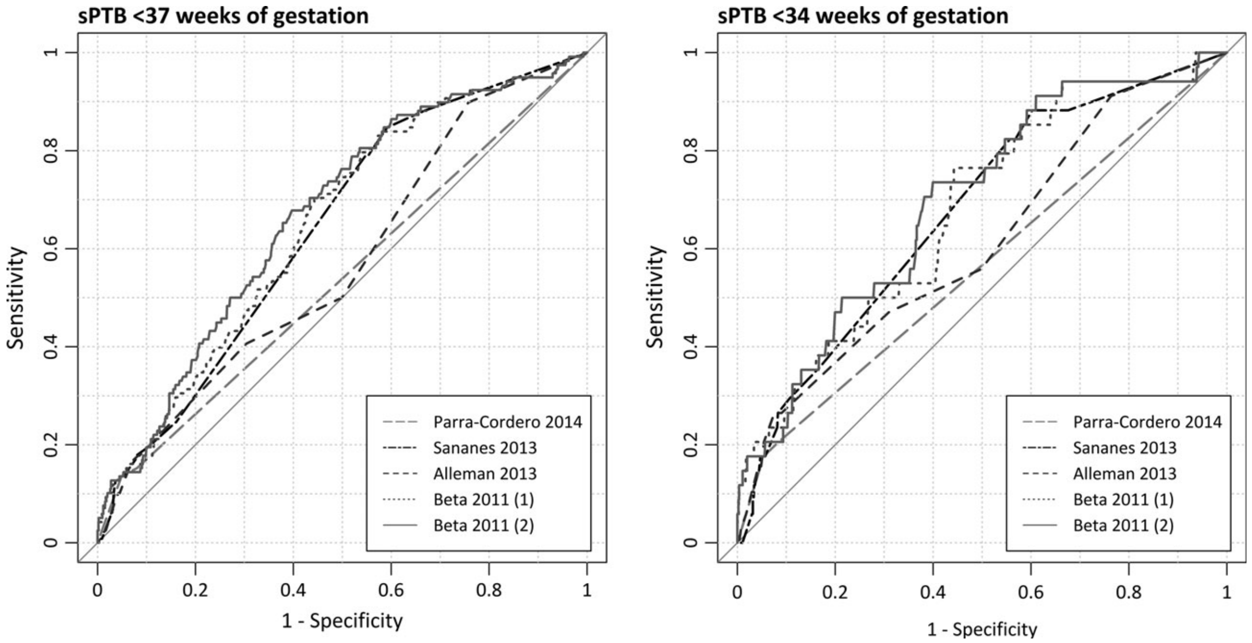

Figure 3. ROC curves of externally validated first trimester prediction models for spontaneous preterm birth (sPTB) <37 weeks of gestation and <34 weeks of gestation.

Moreover, only one study, Beta et al. (29), applied the regression shrinkage technique and only Alleman et al. (27) performed an internal validation by bootstrapping. The methodological limitations mentioned could have been one of the reasons why the reported model performance was not achieved in our validation cohort.

Only Sananes et al. (26) mentioned that they validated their model in another population, but the results were not reported. To our knowledge, no other independent external validation study of prediction models for sPTB exists. External validation is recommended to assess the generalizability to other 'related' populations (24). Our comprehensive independent validation study indicated that all models overestimated performance measures. This illustrates the need for external validation of models before clinical implementation.

Nevertheless, performance measures do not indicate whether a model is clinically useful. Assessment of the clinical utility of the best discriminating model showed a very high false-positive rate at acceptable sensitivity rates. These cut-off points result in a major proportion of nulliparous women being unnecessarily considered to be at

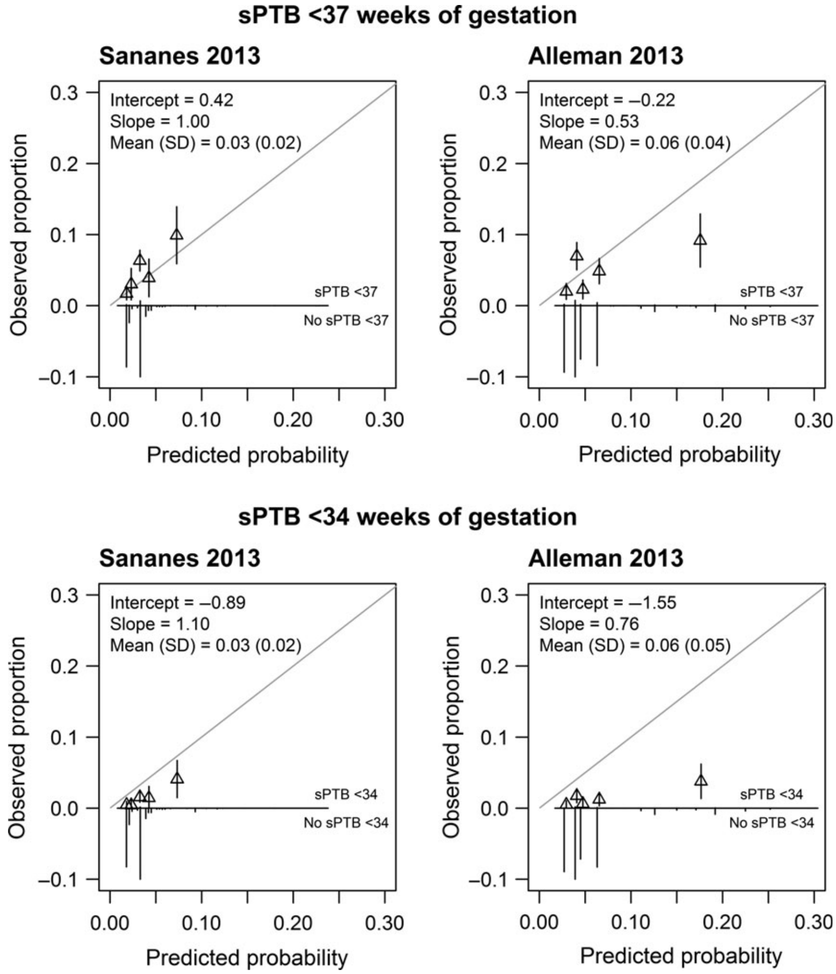

Figure 4. Calibration plots of externally validated first trimester prediction models for spontaneous preterm birth (sPTB) <37 weeks of gestation and <34 weeks of gestation. The gray line is the reference line with intercept = 0 and slope = 1 (perfect calibration). Triangles correspond to grouped predicted risks with 95% confidence intervals (vertical lines).

high risk. Furthermore, for multiparous women the most important predictors are derived from a previous sPTB. In summary, we think that the clinical utility of currently available models is low.

This systematic review demonstrates shortcomings in the quality and performance of existing non-invasive prediction models for sPTB. Improvement of non-invasive models is necessary. The currently available prediction models mainly rely on previous PTB as predicting variable. However, models mainly relying upon a prior event as the discriminative factor do not add much clinical value, since caregivers are already aware that these women are at high risk. Obstetric care would benefit from valid prediction of sPTB in nulliparous women (11).

Future research should focus on the variety of published association studies when selecting candidate predictors. Another important well-known risk factor is cervical surgery (10,33). However, only a minority of women will be identified as high risk by adding this predictor (11). Other routine clinical parameters that may contribute to the prediction of sPTB in nulliparous women are: socioeconomic status, psychological characteristics, family history, medical history, and smoking status (10). Predictive performance of a model might be improved by taking into account biomarkers or ultrasound imaging (for example, cervical length). A few models based on cervical length measurements and biomarkers such as pregnancy-associated plasma protein A (PAPP-A) or alpha-fetoprotein (AFP) have been published (29,34,35). The reported discriminative performance of these models was only slightly better than the performance of models using maternal characteristics alone. We focused in this review on routine clinical

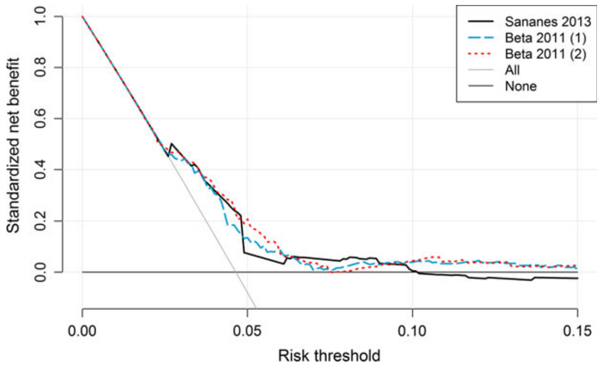

Figure 5. Decision curve analysis of three best performing models for the risk of spontaneous preterm birth <37 weeks of gestation. Decision curve analysis assesses the net benefit (vertical axis; proportion of true-positives and false-positives) of the prediction models over a range of risk thresholds compared with considering all (solid gray line) and no women (horizontal solid black line) to be at high risk for spontaneous preterm birth. [Color figure can be viewed at wileyonlinelibrary.com]

parameters, as these 'specialized' tests are not always routinely performed or readily available in general care, and may generate substantial additional costs (36). Lastly, different modeling methods can be employed as well. In this review, all selected studies used a multiple logistic regression model. Other methods that can be used are machine learning methods using health records, such as tree-based algorithms or neural networks (37,38). However, despite all efforts, sPTB may remain a tough outcome to predict due to its heterogeneous and often unknown causes (2).

Nevertheless, a future model with a moderate performance may still be useful. The trade-off between the benefit of identifying women at high risk and the false-positive rate is important. Using cervical length screening in all women results in the need to screen relatively high numbers of women (11). A non-invasive model combined with a high sensitivity cut-off point would be able to identify women at very low risk of sPTB who could be excluded from cervical length screening, resulting in the need to screen a smaller number of women. Furthermore, such an approach creates the opportunity to identify women at high risk who might benefit from preventive interventions such as progesterone treatment (3–5).

To our knowledge, this is the first systematic review of studies reporting non-invasive prediction models for the risk of sPTB. We had to exclude several published models as three models contained predictors which are not available in the first 16 weeks of pregnancy, for example fetal gender, since this is crucial for early prediction of sPTB. Moreover, three other models did not provide the algorithm, which is essential for independent external validation.

A strength of our study is that we validated all included prediction models in a large independent multicenter prospective cohort of unselected pregnant women. The data were very complete, with a maximum of only 1.2% of missing values. However, although our cohort contained a sufficient number of cases for sPTB <37 weeks of gestation, there were only 34 cases for the secondary outcome sPTB <34 weeks of gestation. An inadequate sample size decreases the precision of external validation measures (22,39).

Our cohort might suffer from treatment bias to a small extent since we did not exclude women who had received treatment such as a cerclage or tocolysis. This may have resulted into the prevention of sPTB and thus an underestimation of model discrimination and calibration (40). One of the selected studies, Alleman et al. (27), explicitly reported exclusion of women undergoing cerclage or tocolysis from their study population (27). Parra-Cordero et al. (28) only excluded women with a history of cerclage (28).

Table 4. Sensitivities, specificities and predictive values at different risk thresholds for model 2 of Beta et al. (29), outcome sPTB <37 weeks of gestation.

Risk threshold[a] (%)	High risk (%)			Sensitivity (%)			Specificity (%)			PPV (%)			NPV(%)		
	All	Nulliparous	Multiparous	All	Nulliparous	Multiparous	All	Nulliparous	Multiparous	All	Nulliparous	Multiparous	All	Nulliparous	Multiparous
2	98.3	100	96.5	99.2	100	97.6	1.8	0	3.5	4.7	6.0	3.3	97.7	100	97.7
3	70.8	98.5	42.5	89.8	100	70.7	30.1	1.6	58.4	5.9	6.1	5.4	98.4	100	98.3
4	51.7	83.3	19.3	76.3	89.6	51.2	49.5	17.1	81.7	6.9	6.5	8.6	97.7	96.3	98.0
5	28.1	41.4	14.6	50.0	49.4	51.2	72.9	59.1	86.7	8.3	7.1	11.5	96.8	94.8	98.1
6	15.2	20.3	9.9	26.3	16.9	43.9	85.3	79.5	91.3	8.1	5.0	14.5	96.0	93.7	98.0
7	10.7	13.9	7.3	19.5	9.1	39.0	89.8	85.7	93.7	8.5	3.9	17.4	95.8	93.7	97.9

NPV, negative predictive value; PPV, positive predictive value.

[a]Predicted risk at or above this level was considered high risk.

Conclusion

This review revealed several reporting and methodological shortcomings of published prediction models for sPTB. Our external validation indicated that none of the models had the ability to predict sPTB adequately in our population. Obstetric care would benefit most from models predicting sPTB accurately among nulliparous women, since most of these women are indicated as low risk in current practice.

Acknowledgments

The Expect Study I could not have been established without the contribution of participating departments of obstetrics and gynecology of hospitals, midwifery practices and maternity care centers in the Province of Limburg: Zuyderland Medical Center Heerlen and Sittard-Geleen, Maastricht University Medical Center, Laurentius Hospital Roermond, Sint Jans Gasthuis Weert, VieCuri Medical Center, midwifery practice Roermond, midwifery practice Nederweert, midwifery practice Weert, midwifery practice Lenie & Chantal, midwifery practice Loes Wijnhoven, midwifery practice De Roerstreek, midwifery practice Bollebuik, midwifery practice Westenberg, midwifery practice Cranendonck, midwifery practice Becca, midwifery practice Born, midwifery practice Geleen, midwifery practice Grevenbicht, midwifery practice Sittard, midwifery practice Sittard-Oost, midwifery practice Stein, midwifery practice Astrea, midwifery practice Horst & Maasdorpen, midwifery practice Reuver-Tegelen, midwifery practice Janneke van Hal, midwifery practice Raijer & Sup, midwifery practice Venlo-Blerick, midwifery practice Venray, midwifery practice Schoffelen-Van Vleuten, midwifery practice Maastricht, midwifery practice Meerssen, midwifery practice Naomi Satijn, midwifery practice Vita, midwifery practice Het Verloskundig Huis, midwifery practice Valkenburg, midwifery practice Parkstad, midwifery practice Lief, midwifery practice Bevalt Beter, midwifery practice 't Bolleke, midwifery practice Natuurlijk bij Jeanny, midwifery practice La Vie, GroenekruisDomicura, Cicogna, ZiNkraamzorg, and maternity center Echt.

Funding

The Expect Study was funded by The Netherlands Organization for Health Research and Development, Pregnancy and Childbirth Program (ZonMw grant 209020007). The funding organization had no role in the design and conduct of the study, analysis or interpretation of data, decision to publish, or preparation of the manuscript.

References

1. Euro-Peristat Project with SCPE and EUROCAT. European Perinatal Health Report. The Health and Care of Pregnant Women and Babies in Europe in 2010. Paris: Euro-Peristat, 2013. Available at: www.europeristat.com
2. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet. 2008;371:75–84.
3. Saigal S, Doyle LW. An overview of mortality and sequelae of preterm birth from infancy to adulthood. Lancet. 2008;371:261–9.
4. Iams JD, Romero R, Culhane JF, Goldenberg RL. Primary, secondary, and tertiary interventions to reduce the morbidity and mortality of preterm birth. Lancet. 2008;371:164–75.
5. Dodd JM, Jones L, Flenady V, Cincotta R, Crowther CA. Prenatal administration of progesterone for preventing preterm birth in women considered to be at risk of preterm birth. Cochrane Database Syst Rev. 2013;(7): CD004947.
6. Romero R, Conde-Agudelo A, Da Fonseca E, O'Brien JM, Cetingoz E, Creasy GW, et al. Vaginal progesterone for preventing preterm birth and adverse perinatal outcomes in singleton gestations with a short cervix: a meta-analysis of individual patient data. Am J Obstet Gynecol. 2018;218:161–80.
7. Gilner J, Biggio J. Management of short cervix during pregnancy: a review. Am J Perinatol. 2016;33:245–52.
8. Alfirevic Z, Stampalija T, Medley N. Cervical stitch (cerclage) for preventing preterm birth in singleton pregnancy. Cochrane Database Syst Rev. 2017;(6): CD008991.
9. Zheng L, Dong J, Dai Y, Zhang Y, Shi L, Wei M, et al. Cervical pessaries for the prevention of preterm birth: a systematic review and meta-analysis. J Matern Fetal Neonatal Med. 2017;1–10. https://10.1080/14767058.2017.1414795
10. Koullali B, Oudijk MA, Nijman TA, Mol BW, Pajkrt E. Risk assessment and management to prevent preterm birth. Semin Fetal Neonatal Med. 2016;21:80–8.
11. Ven J, Os MA, Kazemier BM, Kleinrouweler E, Verhoeven CJ, Miranda E, et al. The capacity of mid-pregnancy cervical length to predict preterm birth in low-risk women: a national cohort study. Acta Obstet Gynecol Scand. 2015;94:1223–34.
12. Esplin MS, Elovitz MA, Iams JD, Parker CB, Wapner RJ, Grobman WA, et al. Predictive accuracy of serial transvaginal cervical lengths and quantitative vaginal fetal fibronectin levels for spontaneous preterm birth among nulliparous women. JAMA. 2017;317:1047–56.
13. Goffinet F. Primary predictors of preterm labour. BJOG. 2005;112:38–47.
14. Honest H, Bachmann LM, Sundaram R, Gupta JK, Kleijnen J, Khan KS. The accuracy of risk scores in

predicting preterm birth – a systematic review. J Obstet Gynaecol. 2004;24:343–59.

15. Kleinrouweler CE, Cheong-See FM, Collins GS, Kwee A, Thangaratinam S, Khan KS, et al. Prognostic models in obstetrics: available, but far from applicable. Am J Obstet Gynecol. 2016;214:79–90. e36.
16. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
17. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. 2014;35:1925–31.
18. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356: i6460.
19. Moons KG, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med. 2014;11: e1001744.
20. Smit HA, Pinart M, Anto JM, Keil T, Bousquet J, Carlsen KH, et al. Childhood asthma prediction models: a systematic review. Lancet Respir Med. 2015;3: 973–84.
21. Meertens LJE, Scheepers HC, De Vries RG, Dirksen CD, Korstjens I, Mulder AL, et al. External validation study of first trimester obstetric prediction models (Expect Study I): research protocol and population characteristics. JMIR Res Protoc. 2017;6:e203.
22. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JD. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol. 2005;58:475–83.
23. Van Buuren S. Flexible imputation of missing data. Boca Raton: CRC Press, 2012.
24. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer Science & Business Media, 2008.
25. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74.
26. Sananes N, Meyer N, Gaudineau A, Aissi G, Boudier E, Fritz G, et al. Prediction of spontaneous preterm delivery in the first trimester of pregnancy. Eur J Obstet Gynecol Reprod Biol. 2013;171:18–22.
27. Alleman BW, Smith AR, Byers HM, Bedell B, Ryckman KK, Murray JC, et al. A proposed method to predict preterm birth using clinical data, standard maternal serum screening, and cholesterol. Am J Obstet Gynecol. 2013;208:472e1–11.
28. Parra-Cordero M, Sepúlveda-Martínez A, Rencoret G, Valdés E, Pedraza D, Muñoz H. Is there a role for cervical assessment and uterine artery Doppler in the first trimester of pregnancy as a screening test for spontaneous preterm delivery? Ultrasound Obstet Gynecol. 2014;43:291–6.
29. Beta J, Akolekar R, Ventura W, Syngelaki A, Nicolaides KH. Prediction of spontaneous preterm delivery from maternal factors, obstetric history and placental perfusion and function at 11–13 weeks. Prenat Diagn. 2011;31:75–83.
30. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMC Med. 2015;13:1.
31. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Basel, Switzerland: Springer, 2015.
32. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127–41.
33. Castanon A, Landy R, Brocklehurst P, Evans H, Peebles D, Singh N, et al. Risk of preterm delivery with increasing depth of excision for cervical intraepithelial neoplasia in England: nested case-control study. BMJ. 2014;349:g6223.
34. van Ravenswaaij R, Tesselaar-van der Goot M, de Wolf S, van Leeuwen-Spruijt M, Visser GH, Schielen PC. First-trimester serum PAPP-A and fbeta-hCG concentrations and other maternal characteristics to establish logistic regression-based predictive rules for adverse pregnancy outcome. Prenat Diagn. 2011;31:50–7.
35. Poon LC, Nekrasova E, Anastassopoulos P, Livanos P, Nicolaides KH. First-trimester maternal serum matrix metalloproteinase-9 (MMP-9) and adverse pregnancy outcome. Prenat Diagn. 2009;29:553–9.
36. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10:e1001381.
37. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2017;38:1805–14.
38. Peissig PL, Santos Costa V, Caldwell MD, Rottscheit C, Berg RL, Mendonca EA, et al. Relational machine learning for electronic health record-driven phenotyping. J Biomed Inform. 2014;52:260–70.
39. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214–26.
40. Pajouheshnia R, Peelen LM, Moons KG, Reitsma JB, Groenwold RH. Accounting for treatment use when validating a prognostic model: a simulation study. BMC Med Res Methodol. 2017;17:103.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Search strategy: Table 1. Framework of systematic research aim according to the CHARMS checklist (19). Figure 1. Flowchart study selection. Search strategy.

Appendix S2. Predictor assessment and model algorithms: Table 1. Definition and assessment predictors included prediction models for spontaneous preterm birth. Table 2. Model algorithms for prediction of spontaneous preterm birth.

Appendix S3. Data extraction and risk of bias assessment: Table 1. Data extraction of included studies according to the CHARMS checklist (19). Table 2. Risk of bias assessment according to the CHARMS checklist and a study of Smit et al. (19,20).

Appendix S4. Characteristics of pregnancies in the observed and imputed validation cohort.

Appendix S5. Baseline characteristics original cohorts and validation cohort.

Appendix S6. Recalibration plots: Figure 1. Calibration plots of recalibrated first trimester prediction models for spontaneous preterm birth (sPTB) <37 weeks of gestation. The gray line is the reference line with intercept = 0 and slope = 1 (perfect calibration). Triangles correspond to grouped predicted risks with 95% confidence intervals (vertical lines). CF, correction factor. Figure 2. Calibration plots of recalibrated first trimester prediction models for spontaneous preterm birth (sPTB) <34 weeks of gestation. The gray line is the reference line with intercept = 0 and slope = 1 (perfect calibration). Triangles correspond to grouped predicted risks with 95% confidence intervals (vertical lines). CF, correction factor.