BJOG An International Journal of Obstetrics and Gynaecology

DOI: 10.1111/1471-0528.15516
www.bjog.org

Epidemiology

External validation and clinical usefulness of first-trimester prediction models for small- and large-for-gestational-age infants: a prospective cohort study

LJE Meertens,[a] LJM Smits,[a] SMJ van Kuijk,[b] R Aardenburg,[c] IMA van Dooren,[d] J Langenveld,[c] IM Zwaan,[e] MEA Spaanderman,[f] HCJ Scheepers[f]

[a] Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, Maastricht, the Netherlands [b] Department of Clinical Epidemiology and Medical Technology Assessment (KEMTA), Maastricht University Medical Centre, Maastricht, the Netherlands [c] Department of Obstetrics and Gynaecology, Zuyderland Medical Centre, Heerlen, the Netherlands [d] Department of Obstetrics and Gynaecology, Sint Jans Gasthuis Weert, Weert, the Netherlands [e] Department of Obstetrics and Gynaecology, Laurentius Hospital, Roermond, the Netherlands [f] Department of Obstetrics and Gynaecology, School for Oncology and Developmental Biology (GROW), Maastricht University Medical Centre, Maastricht, the Netherlands

Correspondence: LJE Meertens, Department of Epidemiology, Care and Public Health Research Institute (CAPHRI), Maastricht University, PO Box 616, 6200 MD, Maastricht, the Netherlands. Email: linda.meertens@maastrichtuniversity.nl

Accepted 12 October 2018. Published Online 17 January 2019.

Objective To assess the external validity of all published first-trimester prediction models based on routinely collected maternal predictors for the risk of small- and large-for-gestational-age (SGA and LGA) infants. Furthermore, the clinical potential of the best-performing models was evaluated.

Design Multicentre prospective cohort.

Setting Thirty-six midwifery practices and six hospitals (in the Netherlands).

Population Pregnant women were recruited at <16 weeks of gestation between 1 July 2013 and 31 December 2015.

Methods Prediction models were systematically selected from the literature. Information on predictors was obtained by a web-based questionnaire. Birthweight centiles were corrected for gestational age, parity, fetal sex, and ethnicity.

Main outcome measures Predictive performance was assessed by means of discrimination (C-statistic) and calibration.

Results The validation cohort consisted of 2582 pregnant women. The outcomes of SGA <10th percentile and LGA >90th percentile occurred in 203 and 224 women, respectively. The C-statistics of the included models ranged from 0.52 to 0.64 for SGA (n = 6), and from 0.60 to 0.69 for LGA (n = 6). All models yielded higher C-statistics for more severe cases of SGA (<5th percentile) and LGA (>95th percentile). Initial calibration showed poor-to-moderate agreement between the predicted probabilities and the observed outcomes, but this improved substantially after recalibration.

Conclusion The clinical relevance of the models is limited because of their moderate predictive performance, and because the definitions of SGA and LGA do not exclude constitutionally small or large infants. As most clinically relevant fetal growth deviations are related to 'vascular' or 'metabolic' factors, models predicting hypertensive disorders and gestational diabetes are likely to be more specific.

Keywords Decision curve analysis, externsal validation, fetal growth, first trimester, large for gestational age, prediction, risk assessment, small for gestational age.

Tweetable abstract The clinical relevance of prediction models for the risk of small- and large-for-gestational-age is limited.

Linked article This article is commented on by J Allotey and S Thangaratinam, p. 485 in this issue. To view this mini commentary visit https://doi.org/10.1111/1471-0528.15564.

Please cite this paper as: Meertens LJE, Smits LJM, van Kuijk SMJ, Aardenburg R, van Dooren IMA, Langenveld J, Zwaan IM, Spaanderman MEA, Scheepers HCJ. External validation and clinical usefulness of first-trimester prediction models for small- and large-for-gestational-age infants: a prospective cohort study. BJOG 2019;126:472–484.

Introduction

Fetal growth deviations are associated with short- and long-term health consequences for both mother and child. Delivering an infant that is large for gestational age (LGA) is associated with trauma to the birth canal, induction of labour, instrumental vaginal delivery, caesarean section, shoulder dystocia, and perinatal asphyxia.[1–5] Infants born small for gestational age (SGA) are at increased risk of perinatal asphyxia, respiratory distress, intubation at term, sepsis, and mortality.[4,6–9] Long-term risks of infants born SGA or LGA are the development of obesity, hypertension, cardiovascular complications, and diabetes later in life.[10–18]

Fetal growth is determined by a complex interplay of genetic factors, uterine conditions, environmental factors, fetal syndromes, hormones, pregnancy complications, and maternal characteristics.[17,19–21] Risk factors for LGA are a high pregestational body mass index (BMI), pre-existing diabetes mellitus, previous LGA, gestational diabetes mellitus (GDM), and a high BMI of the father.[2,22–24] Smoking, short maternal height, chronic hypertension, nulliparity, placental pathology, and intrauterine infections are associated with an increased risk of SGA.[17,25,26] A number of these risk factors are modifiable, but others are not.

The early and correct identification of women at risk would enable personalized follow-up management, which could help to avoid adverse perinatal outcomes. Prediction modelling combines risk factors into a single model that takes into account the risk-dependent weight of each factor and possible interrelations.[27,28] Several prediction models based on maternal characteristics, biomarkers, and biophysical tests have been developed for the risk of SGA or LGA, showing promising discriminative performance in separating fetal growth deviations from normal growth. Biomarkers and biophysical tests may improve the accuracy of the model beyond using maternal characteristics alone. Published studies show only a limited contribution of these factors to improved discriminative performance, however.[29–46] Moreover, most of these more complex predictors are relatively expensive, not readily available in general antenatal settings, and are possibly inconvenient for pregnant women.[28] To our knowledge, no external validation studies of prediction models for SGA or LGA have been published so far. External validation is a crucial step before implementing a model in clinical practice, by evaluating the performance using data that were not used to develop the model.[47]

In this study, an overview of all published prediction models for the risk of SGA or LGA based on maternal characteristics and standard antenatal measurements (i.e. blood pressure) is provided. We validated the selected models in an independent Dutch prospective cohort consisting of 2582 pregnant women. Furthermore, we evaluated the clinical potential of the best-performing models.

Methods

Selection of prediction models

We systematically searched PubMed to select all published early prediction models, based on routinely collected parameters and applicable in the first trimester of pregnancy, for the risk of SGA or LGA. The searches were performed in April 2013, before the development of the study questionnaires, and were updated until 22 June 2017. The search strategies and selection criteria have been published elsewhere.[48]

Validation cohort

We performed a multicentre prospective cohort study in the south-eastern part of the Netherlands (Expect Study I). The primary objective of this study was to validate published first-trimester prediction models for several adverse pregnancy outcomes. Six hospitals and 36 midwifery practices recruited pregnant women aged ≥18 years old at <16 weeks of gestation between 1 July 2013 and 1 January 2015, with follow-up continuing until 31 December 2015. Eligible pregnant women were invited to complete two web-based questionnaires (paper-based questionnaires were available, upon request): one before 16 weeks of gestation (pregnancy questionnaire) and one 6 weeks after the due date (postpartum questionnaire). Medical records and discharge letters were requested from health care providers. Pregnancies ending in miscarriage (<16 weeks of gestation), terminations of pregnancy before 24 weeks of gestation, and women lost-to-follow-up were excluded. For this study, we also excluded multiple pregnancies and women who delivered between 16^{+0} and 25^{+0} weeks of gestation, as the customised birthweight curves are only available from 25 weeks of gestation onwards.[49] A detailed description of Expect Study I has been published in full elsewhere.[48] Patients were involved in the development of the recruitment process and the study questionnaires. The design, results, and conclusions of this pilot study are described in the published study protocol.[48] The study was funded by The Netherlands Organization for Health Research and Development (ZonMw grant 209020007).

Assessment of predictors and outcomes

The predictors in the included prediction models were assessed by means of the pregnancy questionnaire. Blood pressure was measured according to routine antenatal care and self-reported in the pregnancy questionnaire. We used the same definitions as published in the original articles (Appendix S1; Tables 1 and 2).

Table 1. Baseline characteristics of the validation cohort (Expect Study I)

Characteristics	Missing values n (%)	Observed validation cohort (Expect Study I)*				
		Overall (*n* = 2582)	SGA <10th percentile (*n* = 203)	No SGA (*n* = 2379)	LGA >90th percentile (*n* = 224)	No LGA (*n* = 2358)
Age, years	0 (0.0)	30.2 (3.9)	30.0 (4.4)	30.2 (3.9)	30.1 (3.8)	30.2 (3.9)
Ethnicity	0 (0.0)					
White		2503 (96.9)	197 (97.0)	2306 (96.9)	218 (97.3)	2285 (96.9)
Afro-Caribbean		2 (0.1)	1 (0.5)	1 (0.0)	0 (0.0)	2 (0.1)
South Asian		4 (0.2)	0 (0.0)	4 (0.2)	0 (0.0)	4 (0.2)
East Asian		16 (0.6)	1 (0.5)	15 (0.6)	0 (0.0)	16 (0.7)
Other Asian		0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
Hispanic		11 (0.4)	0 (0.0)	11 (0.5)	1 (0.4)	10 (0.4)
Mixed		46 (1.8)	4 (2.0)	42 (1.8)	5 (2.2)	41 (1.7)
Tertiary education	3 (0.1)	1403 (54.3)	90 (44.3)	1313 (55.2)	134 (59.8)	1269 (53.8)
Height, cm	3 (0.1)	168.8 (6.4)	166.2 (6.2)	169.0 (6.4)	171.7 (6.4)	168.5 (6.3)
Weight, kg	4 (0.2)	68.9 (13.0)	65.2 (11.7)	69.2 (13.1)	75.4 (13.7)	68.3 (12.8)
Body mass index, kg/m^2	4 (0.2)	24.2 (4.3)	23.5 (3.7)	24.2 (4.3)	25.6 (4.7)	24.0 (4.2)
<18.5		87 (3.4)	8 (3.9)	79 (3.3)	2 (0.9)	85 (3.6)
18.5–24.9		1645 (63.7)	136 (67.0)	1509 (63.4)	120 (53.6)	1525 (64.7)
25.0–29.9		576 (22.3)	46 (22.7)	530 (22.3)	61 (27.2)	515 (21.8)
≥30.0		270 (10.4)	13 (6.4)	257 (10.8)	41 (18.3)	229 (9.7)
Smoking	1 (0.0)					
Ever <16 weeks of gestation		312 (12.1)	54 (26.6)	258 (10.8)	25 (11.2)	287 (12.2)
Current (at completion questionnaire)		152 (5.9)	35 (17.2)	117 (4.9)	7 (3.1)	145 (6.1)
Diabetes mellitus	0 (0.0)	11 (0.4)	0 (0.0)	11 (0.5)	4 (1.8)	7 (0.3)
Type 1		9 (0.3)		9 (0.4)	3 (1.3)	6 (0.3)
Type 2		1 (0.0)		1 (0.0)	1 (0.4)	0 (0.0)
Other	0 (0.0)	1 (0.0)		1 (0.0)	0 (0.0)	1 (0.0)
History of chronic hypertension	0 (0.0)	27 (1.0)	0 (0.0)	27 (1.1)	4 (1.8)	23 (1.0)
Folate use at completion questionnaire	3 (0.1)	2198 (85.1)	165 (81.3)	2033 (85.5)	189 (84.4)	2009 (85.2)
Parity	0 (0.0)					
Nulliparous		1311 (50.8)	103 (50.7)	1208 (50.8)	121 (54.0)	1190 (50.5)
Primiparous		1015 (39.3)	74 (36.5)	941 (39.6)	74 (33.0)	941 (39.9)
Multiparous		256 (9.9)	26 (12.8)	230 (9.6)	29 (12.9)	227 (9.6)
Conception	0 (0.0)					
Spontaneous		2412 (93.4)	187 (92.1)	2225 (93.5)	207 (92.4)	2205 (93.5)
Ovulation induction		92 (3.6)	11 (5.4)	81 (3.4)	10 (4.5)	82 (3.5)
IVF/ICSI		78 (3.0)	5 (2.5)	73 (3.1)	7 (3.1)	71 (3.0)
Interpregnancy interval, months	11 (0.4)	29.0 (24.2)	32.0 (29.5)	28.7 (23.7)	25.8 (17.9)	29.3 (24.7)
History of SGA						
<5th percentile	51 (2.0)	42 (1.6)	8 (3.9)	34 (1.4)	1 (0.4)	41 (1.7)
<10th percentile	51 (2.0)	106 (4.1)	18 (8.9)	88 (3.7)	1 (0.4)	105 (4.5)
History of LGA						
>90th percentile	51 (2.0)	167 (6.5)	1 (0.5)	166 (7.0)	44 (19.6)	123 (5.2)
>95th percentile	51 (2.0)	89 (3.4)	0 (0.0)	89 (3.7)	30 (13.4)	59 (2.5)
Birthweight *z*-score of previous pregnancy	49 (1.9)	0.15 (1.0)	−0.55 (0.7)	0.21 (1.0)	1.13 (1.0)	0.07 (0.9)
History of pregnancy induced hypertension	18 (0.7)	114 (4.4)	10 (4.9)	104 (4.4)	9 (4.0)	105 (4.5)
History of pre-eclampsia	18 (0.7)	71 (2.7)	5 (2.5)	66 (2.8)	3 (1.3)	68 (2.9)
History of gestational diabetes mellitus	19 (0.7)	14 (0.5)	1 (0.5)	13 (0.5)	5 (2.2)	9 (0.4)
Systolic blood pressure, mmHg	257 (10.0)	114.4 (12.4)	115.0 (13.0)	114.3 (12.4)	116.3 (11.7)	114.2 (12.5)
Diastolic blood pressure, mmHg	267 (10.3)	67.6 (8.5)	67.6 (8.2)	67.6 (8.6)	68.6 (8.3)	67.5 (8.5)
Mean arterial pressure, mmHg	267 (10.3)	83.2 (8.8)	83.4 (8.5)	83.2 (8.8)	84.5 (8.4)	83.1 (8.8)

Abbreviations: ICSI, intracytoplasmic sperm injection; IVF, *in vitro* fertilization; LGA, large-for-gestational-age; SGA, small for gestational age.
*Original data (not imputed) presented as mean (SD) or absolute number (%).

Table 2. Discriminative performance of included prediction models for small-for-gestational-age infants

Study (author, year)	AUROC (95% CI) Original publication	AUROC (95% CI) Validation cohort SGA <5th percentile (*n* = 104)	AUROC (95% CI) Validation cohort SGA <10th percentile (*n* = 203)	AUROC (95% CI) Validation cohort, nulliparous (*n* = 1311) SGA <10th percentile (*n* = 103)	AUROC (95% CI) Validation cohort, multiparous (*n* = 1271) SGA <10th percentile (*n* = 100)
González González (2017)[33]	0.615 (0.571–0.658) Internal validation: 0.594 (NR)	0.60 (0.55–0.65)	0.57 (0.53–0.61)	0.56 (0.52–0.59)	0.57 (0.53–0.61)
MacDonald-Wallis (2015)[57]	0.70 (0.68–0.71)	0.67 (0.61–0.72)	0.63 (0.59–0.67)	0.66 (0.60–0.71)	0.67 (0.61–0.72)
Boucoiran (2013)[29]	0.62 (0.58–0.66)	0.63 (0.56–0.59)	0.64 (0.60–0.68)	0.64 (0.59–0.70)	0.64 (0.58–0.70)
Syngelaki (2011)[58]	NR	0.59 (0.54–0.65)	0.58 (0.54–0.62)	0.56 (0.50–0.61)	0.61 (0.55–0.67)
Poon (2011)[42]	0.719 (0.706–0.732)	0.52 (0.46–0.57)	0.52 (0.48–0.56)	0.52 (0.47–0.58)	0.52 (0.46–0.58)
Seed (2011)[59]	0.65 (NR) Internal validation: 0.57 (NR)	0.55 (0.50–0.61)	0.54 (0.50–0.58)	0.50 (0.45–0.56)	0.57 (0.51–0.63)

AUROC, area under the receiving operating characteristic curve; CI, confidence interval; NR, not reported; SGA, small for gestational age.

The outcomes of SGA and LGA were defined as infants with a birthweight below the tenth percentile or above the 90th percentile, respectively, corrected for gestational age, ethnicity, gender, and parity.[49] We also evaluated the performance of the models for SGA and LGA using the cut-off values of the fifth percentile and the 95th percentile, respectively. Birthweight was obtained from the medical records. Data from the postpartum questionnaire were used in the case of a missing birthweight (n = 1) or absence of the medical record (n = 16).

Statistical analysis

There is no generally accepted rule for the required sample size for external validation of prediction models. We followed Vergouwe et al.,[50] who recommend a minimum of 100 events and 100 non-events.

The baseline characteristics of the validation cohort were described as mean ± standard deviation (SD) for continuous variables and as an absolute value with percentage for categorical variables. Stochastic regression imputation was used to enter missing predictor variables, with predictive mean matching used as the imputation model.[51] We also evaluated the similarity of the validation cohort to the derivation cohorts.

We computed the individual probabilities for the risk of SGA or LGA using the original prediction algorithms (Appendix S1; Tables 3 and 4). The predictive performance of each model was assessed by means of discrimination and calibration. Discriminative performance, i.e. the ability of the model to distinguish between women who will have the outcome and those who will not, was quantified as the area under the receiver operating characteristic curve (AUROC) with 95% confidence interval (95% CI).[47] A subgroup analysis was performed for nulliparous women, as a history of SGA or LGA is a strong predictor for recurrent SGA or LGA, respectively. Calibration is a measure of the agreement between the predicted probabilities of the model and the actual outcomes.[47] We assessed calibration graphically by calibration plots, in which women were divided into groups of equal size (up to ten) with similar predicted risks, and computed calibration-in-the-large and the calibration slope. Calibration-in-the-large (intercept) indicates whether predictions are systematically too low (intercept >0 | slope = 1) or too high (intercept <0 | slope = 1) by comparing the mean predicted risk with the observed proportion of cases.[47] The slope refers to the average strength of the predictor effects (overfitting, <1; underfitting, >1). Calibration plots that indicate perfect agreement have an intercept of 0 and a slope of 1 (45° line).[47] We recalibrated the models by adjusting the intercept and slope using the linear predictor as the only covariate. This recalibration method has no influence on the discriminative performance.[52]

Lastly, we evaluated the clinical potential of the best-performing models by means of decision curve analysis. Decision curve analysis provides insight into the net benefit of the prediction model over a range of risk thresholds compared with the scenarios that all ('treat all') or no ('treat none') women are at high risk of the outcome.[53] The net benefit of a model can be clinically interpreted as the net increase in the proportion of appropriately treated patients

Table 3. Discriminative performance of included prediction models for large-for-gestational-age infants

Study (author, year)	AUROC (95% CI) Original publication	AUROC (95% CI) Validation cohort LGA >95th percentile (*n* = 105)	AUROC (95% CI) Validation cohort LGA >90th percentile (*n* = 224)	AUROC (95% CI) Validation cohort, nulliparous (*n* = 1311) LGA >90th percentile (*n* = 121)	AUROC (95% CI) Validation cohort, multiparous (*n* = 1271) LGA >90th percentile (*n* = 103)
Frick (2016)[31]	NR	0.74 (0.70–0.79)	0.69 (0.66–0.72)	0.66 (0.61–0.71)	0.80 (0.76–0.84)
González González (2013)[32]	0.680 (0.659–0.700)	0.67 (0.62–0.72)	0.64 (0.60–0.68)	0.67 (0.62–0.72)	0.70 (0.64–0.74)
Plasencia (2012)[40]	0.705 (0.684–0.725)	0.64 (0.59–0.70)	0.62 (0.58–0.65)	0.66 (0.61–0.72)	0.70 (0.65–0.75)
Syngelaki (2011)[58]	NR	0.64 (0.58–0.70)	0.60 (0.56–0.64)	0.58 (0.53–0.63)	0.71 (0.66–0.77)
Nanda (2011)[37]	0.722 (0.710–0.735)	0.73 (0.68–0.78)	0.68 (0.64–0.71)	0.64 (0.59–0.69)	0.73 (0.69–0.78)
Poon (2011)[42]	0.715 (0.710–0.719)	0.73 (0.68–0.78)	0.68 (0.64–0.71)	0.64 (0.59–0.69)	0.74 (0.70–0.79)

AUROC, area under the receiving operating characteristic curve; CI, confidence interval; LGA, large for gestational age; NR, not reported.

Table 4. Performance measures at different risk thresholds for the recalibrated model from Boucoiran (2013),[29] predicting the risk of small-for-gestational-age infants

Risk threshold*, %	High risk, % (*n*/*n*)	Sensitivity, % (*n*/*n*)	Specificity, % (*n*/*n*)	PPV, % (*n*/*n*)	NPV, % (*n*/*n*)
2	99.6 (2571/2582)	100 (203/203)	0.46 (11/2379)	7.9 (203/2571)	100 (11/11)
4	92.1 (2377/2582)	97.0 (197/203)	8.4 (199/2379)	8.3 (197/2377)	97.1 (199/205)
6	70.1 (1811/2582)	82.3 (167/203)	30.9 (735/2379)	9.2 (167/1811)	95.3 (735/771)
8	32.1 (829/2582)	52.7 (107/203)	69.7 (1657/2379)	12.9 (107/829)	94.5 (1657/1753)
10	9.7 (250/2582)	25.1 (51/203)	91.6 (2180/2379)	20.4 (51/250)	93.5 (2180/2332)
12	5.9 (153/2582)	17.2 (35/203)	95.0 (2261/2379)	22.9 (35/153)	93.1 (2261/2429)
14	5.5 (143/2582)	15.8 (32/203)	95.3 (2268/2379)	22.4 (32/143)	93.0 (2268/2439)

NPV, negative predictive value; PPV, positive predictive value.
*Predicted risk at or above this level was considered as high risk.

or as the net decrease in the proportion of patients treated unnecessarily (proportion of true positives and false positives).[54] Next, we calculated sensitivity, specificity, and positive and negative predictive values at certain risk thresholds for the models with the highest overall net benefit.

Statistical analyses were performed with R 3.4.1 (using the packages MICE, RMS, PROC, and DECISIONCURVE).

Results

Selection of prediction models

The search strategies identified 1522 and 334 articles for SGA and LGA, respectively. Fifteen articles fulfilled the eligibility criteria for the outcome SGA. The cross-checking of citation lists yielded three additional articles. We excluded ten articles for the following reasons: algorithm not available (n = 8),[35,38–40,43,45,46,55] predictors not applicable in a high-income country (n = 1),[56] and model already published in another included article (n = 1).[34] The eight eligible articles described nine prediction models aimed at predicting any SGA (n = 6),[29,33,44,57–59] preterm SGA (n = 2),[60,61] and late SGA (n = 1).[60] For the outcome LGA, we selected nine eligible articles all describing a prediction model for any LGA.[31,32,37,39–42,58,62] No additional articles were found by cross-checking references. Three articles were excluded as the algorithm was not available.[39,40,62]

We only validated models predicting the risk of any SGA or any LGA, as the number of preterm SGA infants was too low in our validation cohort (n = 6 at <37 weeks of gestation). None of the models were used in antenatal care during the study period.

The characteristics of the included models for SGA (n = 6),[29,33,44,57–59] or for LGA (n = 6),[31,32,37,41,42,58] are summarized in Tables S1 and S2, respectively. The models for SGA were published by five different research groups

from the UK, Canada, and Spain between 2011 and 2017. Four models originally defined SGA as an infant with a birthweight below the tenth percentile, and two models used a cut-off value of below the fifth percentile. Two studies used the same study data to determine the birthweight centiles. The other studies used national charts ($n = 1$), their own study data ($n = 1$), or a customised birthweight calculator ($n = 1$). Three different research groups from the UK and Spain published models predicting the risk of LGA between 2011 and 2016. Four models defined LGA as a birthweight above the 95th percentile, based on the same study data, and two models used a birthweight above the 90th percentile, based on national charts.

Validation cohort

We included 2582 women in the validation cohort (Figure S1). The outcome of SGA with a birthweight below the tenth percentile occurred in 203 women (7.9%). Six SGA infants were born prematurely (<37 weeks of gestation) (3.0%), and 14 SGA infants were born to mothers whose pregnancy was complicated by a hypertensive pregnancy disorder (6.9%). Of the 224 infants who were LGA, with birthweights above the 90th percentile (8.7%), 20 were born to mothers with GDM (8.9%). Table 1 shows the characteristics of the overall cohort, and for SGA, non-SGA, LGA, and non-LGA groups in the observed data. The characteristics of the imputed validation cohort were generally comparable with those of the observed data (Table S3). We also compared the characteristics of the validation cohort with the derivation cohorts (Appendix S2; Tables 1 and 2). In contrast to most derivation cohorts, our validation cohort had a low prevalence of non-white ethnicity and smoking during pregnancy. The average height and weight of the women was higher compared with all other development cohorts, but the mean BMI was similar. The occurrence of the outcome SGA was considerably higher in the derivation cohorts of Seed et al. (high-risk women), [59] and of González González et al.[33] The prevalence of LGA was comparable between the development cohorts and our validation cohort. Compared with all other derivation cohorts, nulliparous women in our cohort delivered LGA infants more often than not. Syngelaki et al.[58] neither reported predictor characteristics nor reported the numbers of SGA and LGA infants.

Predictive performances

Table 2 shows the AUROCs for the prediction of SGA. The discriminative performance decreased considerably for all models compared with the development cohorts, but decreased most for the models with the highest AUROCs. The model of Boucoiran et al. retained the highest AUROC (0.64, 95% CI 0.60–0.68).[29] All models demonstrated a higher ability to predict the risk of SGA below the fifth percentile compared with SGA below the tenth percentile, with AUROCs of up to 0.67; however, the 95% CIs were wider because of the lower sample size, with lower limits close to 0.50. Subgroup analysis showed no difference in the discriminative ability of the models between nulliparous and multiparous women. The ROC curves are displayed in Figure S2. The three models for which the full algorithm was provided showed poor calibration (Figure 1). The recalibration of all models led to better agreement between the predicted and the observed risks of most models (Figure S3). The model of MacDonald-Wallis et al.[57] showed the closest fit to the perfect calibration line. The predicted risks of all models were closely clustered around the overall risk.

The discriminative performances of the prediction models for LGA are presented in Table 3. Although the AUROCs also decreased for all models after external validation, three models showed moderate discriminative ability (AUROCs 0.68–0.69). The model of Frick et al. showed the highest discriminative performance with an AUROC of 0.69 (95% CI 0.66–0.72) and 0.74 (95% CI 0.70–0.79) for LGA >90th percentile and >95th percentile, respectively.[31] All models showed a higher discriminative ability for LGA >95th percentile compared with LGA >90th percentile. In contrast to the outcome SGA, most models for LGA were also originally developed to predict the 5% of most extreme birthweight deviations (>95th percentile). Figure S4 presents the ROC curves. Subgroup analysis showed better discriminative performance among multiparous women, with the highest AUROC (0.80) for the model of Frick et al.[31] Performance among nulliparous women was slightly lower than in the total group (AUROC up to 0.67). The three fully available algorithms for LGA showed better calibration as compared with models for SGA (Figure 2). All models overestimated the probabilities on average (intercept < 0) and showed an overfitting of the predictor effects (slope < 1, low predictions too low, and high predictions too high). Recalibration of all models considerably improved the agreement between the predicted probabilities and the observed outcomes for almost all models (Figure S5).

Clinical usefulness

Decision curve analysis of the two best-performing models for SGA revealed a positive net benefit compared with classifying all or no women as being at high risk of SGA for a risk threshold of 4–22% (Figure 3); however, the overall net benefit remained low throughout this range. This low clinical usefulness is also demonstrated in Table 4. Choosing a low cut-off is associated with a high sensitivity, so the risk of missing women with the outcome is minimized (low rate of false negatives). High sensitivity leads to a large proportion of women unnecessarily indicated as being at high risk, however. A higher risk threshold ensures a

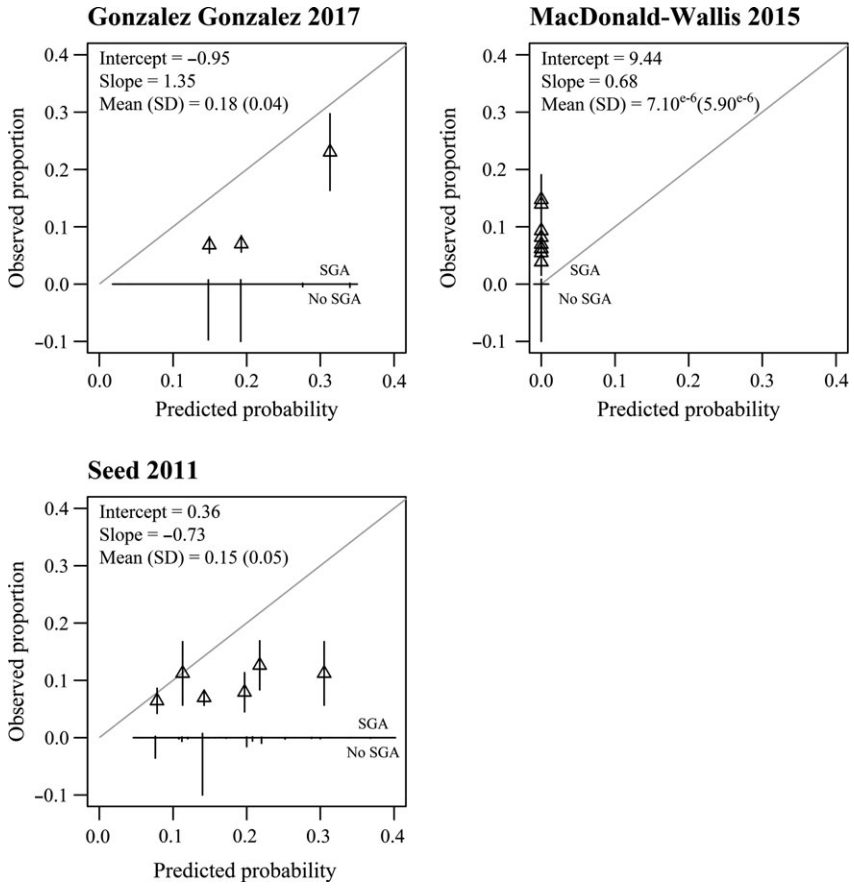

Figure 1. Calibration plots of externally validated first-trimester prediction models for small-for-gestational-age infants with birthweights below the tenth percentile. The grey line is the reference line with intercept = 0 and slope = 1 (perfect calibration). Triangles correspond to grouped predicted risks with 95% confidence intervals (vertical lines).

high specificity (low rate of false positives), but only a small proportion of women with the outcome will be detected.

Figure 4 shows the net benefit of the three best discriminative models for the risk of LGA. The models were beneficial compared with classifying all or no women as being at high risk over a threshold range of 1–40%. The three curves differed only slightly. Table 5 presents the performance measures at different risk thresholds.

Discussion

Main findings

Six early non-invasive prediction models for the risk of SGA as well as six models for the outcome LGA were selected from the literature, some of which showed promising original discriminative performance (AUROC up to 0.72). We validated these models in an independent prospective cohort of 2582 women. The discriminative performance decreased for all models, especially those predicting SGA. All models showed better discriminative ability for predicting the more severe cases of SGA and LGA, which are also associated with a higher risk of adverse outcomes.[31,63] Calibration was poor for the prediction models for SGA. The models predicting the risk of LGA all overestimated the risk in our population. Recalibration provided better agreement between the predicted risks and the actual risks for most models. The predictive performance is usually lower in other populations, even when a similar population as the one in which the model was developed is being used.[47] The studies used different sources to define the birthweight centile (i.e. customised and population-based charts) that may have contributed to the different performance in our population. Evaluation of the predictive performance of all available models in one independent cohort allowed for a fair comparison.

Interpretation

The validation of promising prediction models is essential in order to gain insight into their robustness across other populations. Only two of the selected models were internally validated and their performance stayed fairly stable after

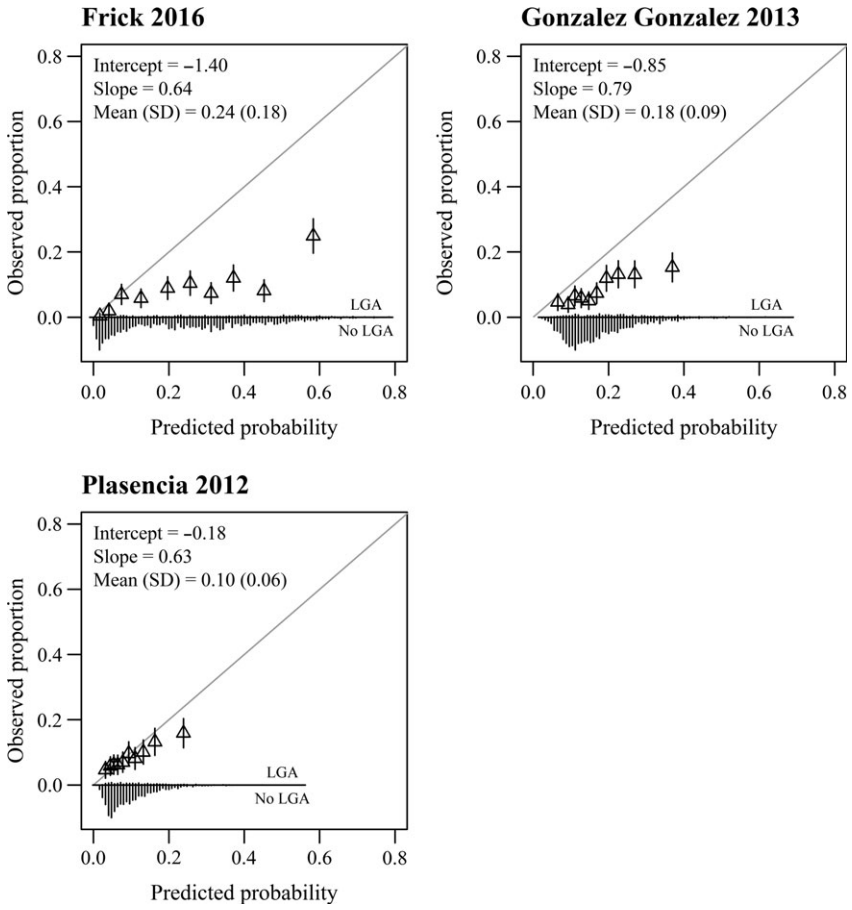

Figure 2. Calibration plots of externally validated first-trimester prediction models for large-for-gestational-age infants with birthweights above the 90th percentile. The grey line is the reference line with intercept = 0 and slope = 1 (perfect calibration). Triangles correspond to grouped predicted risks with 95% confidence intervals (vertical lines).

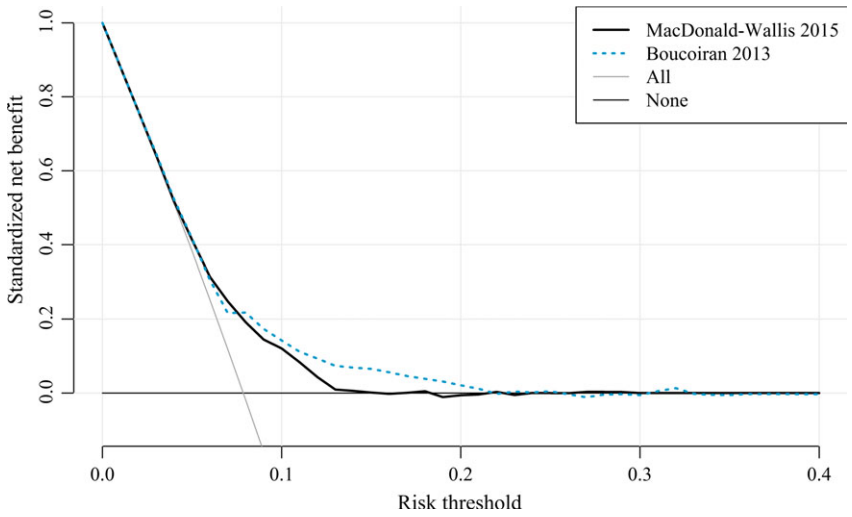

Figure 3. Decision curve analysis of the two best-performing models for the risk of small-for-gestational-age infants with birthweights below the tenth percentile. The solid grey line is the net benefit when considering all women as being at high risk, and the horizontal black line is the net benefit when considering no women being at high risk.

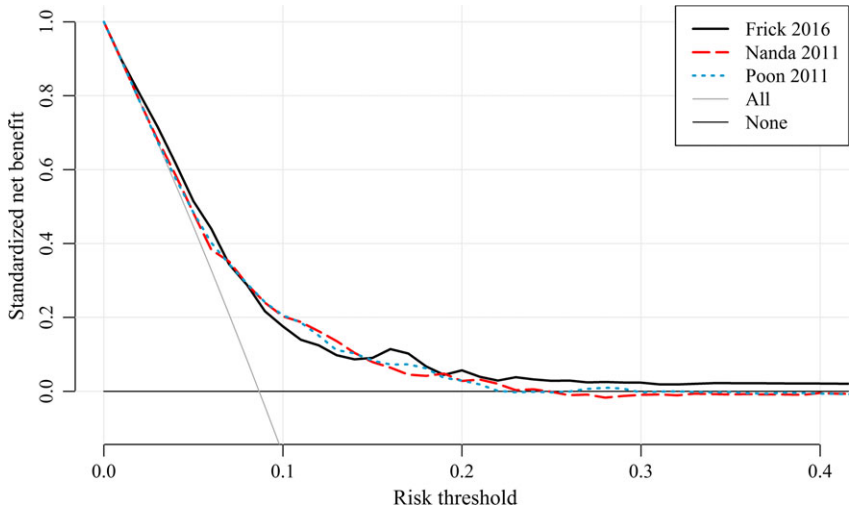

Figure 4. Decision curve analysis of the three best-performing models for the risk of large-for-gestational-age infants with birthweights above the 90th percentile. The solid grey line is the net benefit when considering all women as being at high risk, and the horizontal black line is the net benefit when considering no women being at high risk.

Table 5. Performance measures at different risk thresholds for recalibrated model from Frick et al.,[31] predicting the risk of large-for-gestational-age

Risk threshold*, %	High risk, % (*n*/*n*)	Sensitivity, % (*n*/*n*)	Specificity, % (*n*/*n*)	PPV, % (*n*/*n*)	NPV, % (*n*/*n*)
1	98.5 (2541/2582)	100 (224/224)	1.7 (41/2358)	8.8 (224/2541)	100 (41/41)
2	89.7 (2424/2582)	99.6 (223/224)	11.3 (266/2358)	9.6 (223/2424)	99.6 (266/267)
4	72.7 (1877/2582)	92.9 (208/224)	29.2 (689/2358)	11.1 (208/1877)	97.7 (689/705)
8	50.7 (1309/2582)	73.2 (164/224)	51.4 (1213/2358)	12.5 (154/1309)	95.3 (1213/1273)
10	37.8 (976/2582)	59.4 (133/224)	64.2 (1515/2358)	13.6 (133/976)	94.3 (1515/1606)
14	17.0 (438/2582)	34.8 (78/224)	84.7 (1998/2358)	17.8 (78/438)	93.2 (1998/2144)
18	5.7 (148/2582)	17.4 (39/224)	95.4 (2249/2358)	26.4 (39/148)	92.4 (2249/2434)
20	2.7 (69/2582)	10.7 (24/224)	98.1 (2313/2358)	34.8 (24/69)	92.0 (2313/2513)

NPV, negative predictive value; PPV, positive predictive value.
*Predicted risk at or above this level was considered as high risk.

external validation.[33,59] MacDonald-Wallis et al. validated their developed model in another cohort from the same country.[57] To our knowledge, no independent or other external validation studies have been published on prediction models for SGA or LGA. Validating prediction models in an independent population provides insight into the generalisability of the model, an essential element before clinical application can be considered.[28] Further research should focus on validation and the updating of existing models instead of investing energy in developing yet another prediction model similar to those already available.[64]

Predictive performance measures of a prediction model do not coincide with the usefulness of the model in clinical practice. Decision curve analysis and performance parameters at different risk thresholds give a first impression of the clinical utility. A prediction tool should ideally separate individuals who will develop the disease from those who will not. But in fact, there is a trade-off between sensitivity and specificity. Evaluation of the clinical utility is therefore dependent on several other factors, such as the severity of the health consequences related to fetal growth deviations and the availability of effective follow-up management, to prevent either the development of fetal growth deviations (primary) or the related adverse effects (secondary).

The heterogeneous aetiology of fetal growth deviations makes prediction difficult.[24,65] Infants who are constitutionally SGA or LGA are less likely to develop adverse outcomes and also less likely to benefit from interventions.[65,66] A subset of possibly clinically relevant SGA and LGA frequently has a 'metabolic' (i.e. high BMI or GDM) or 'vascular' (i.e., hypertensive disorder) origin. The predictors in the included models for SGA and LGA overlap

considerably with those of models predicting hypertensive pregnancy disorder and GDM, respectively.[67,68] Although most SGA and LGA infants are born to mothers without a hypertensive pregnancy disorder or GDM, respectively, the conditions share common pathophysiological aspects.[24,65]

Regarding primary prevention strategies, recent meta-analyses have demonstrated that aspirin modestly reduces the risk of delivering an SGA infant in women at high risk, with most benefit derived by starting treatment before 16 weeks of gestation and using a dose of ≥100 mg (risk ratio 0.56–0.76).[69,70] Patient selection of those at increased risk was primarily based on an increased risk of developing a hypertensive pregnancy disorder rather than delivering an SGA infant.[71] Currently, there are no effective interventions for the primary prevention of LGA available, except for the treatment of women with GDM that indirectly lowers the risk of LGA, such as diet.[24] In summary, the application of the currently available prediction models for the risk of SGA or LGA, in settings in which models for the identification of 'vascular' (pre-eclampsia)- and 'metabolic' (GDM)-related complications are already applied, are not likely to result in additional benefit regarding the overlap of predictors and preventive interventions.

The identification of women at risk may also allow for the secondary prevention of adverse effects related to SGA and LGA. Antenatal detection of infants born SGA and delivery at the appropriate time may reduce the risk of severe morbidity and mortality.[72–75] Induction of labour at or near term for pregnancies suspected to deliver a LGA infant results in a lower mean birthweight, and fewer birth fractures and shoulder dystocia.[76,77] In most clinical settings, ultrasound fetal biometry is the current method for the prediction of birthweight. Based on the decision curve analysis, the use of prediction models to select women for ultrasound fetal biometry will probably not be any more beneficial compared with providing this for all women. Moreover, should ultrasound fetal biometry be restricted to high-risk women, it is again clinically relevant that the model selects the pathological fetal growth deviations. Another important aspect is that even infants who do not meet the criteria for SGA or LGA can have a pathological growth pattern, such as asymmetrical growth or a declining or accelerated growth pattern.[66] These pathological growth patterns are also likely to be related to 'vascular' and 'metabolic' complicated pregnancies, and serial ultrasound fetal biometry is needed for detection. In conclusion, models that would predict pathological fetal growth deviations are more likely to improve clinical outcomes than models predicting SGA or LGA.

Strengths and limitations

We externally validated all published non-invasive prediction models in an independent population. The multicentre prospective study design, with no strict inclusion criteria, ensured the sample was as heterogeneous as possible. Our data contained a low level of missing data (<1% for most predictors) and out-of-range values, as we incorporated validation checks in the web-based questionnaires. Missing data were handled by the use of imputation in order to prevent biased results. Nevertheless, a substantial number of blood pressure measurements were missing (10%), most likely because of the self-reporting of measurements in the web-based questionnaire. Only two models contained a predictor based on blood pressure measurement. Another limitation to be mentioned is that we had to exclude women who delivered between 16^{+0} and 24^{+6} weeks of gestation ($n = 8$), as the Dutch population-based reference curves for birthweight centiles are available from 25 weeks of gestation onwards.[49] Lastly, we had to exclude two prediction models in the selection process, as we did not dispose of routine blood parameters (random glucose, rhesus group) and ultrasound measurements (crown–rump length).[36,78]

Conclusion

The clinical relevance of prediction models for SGA and LGA can be questioned, both for the moderate predictive performance and the heterogeneous aetiology of fetal growth deviations. It is important to distinguish between constitutional and pathological fetal growth deviations to improve clinical outcomes. Not much additional clinical benefit is expected of the current prediction models for SGA and LGA over models that predict pre-eclampsia and GDM, due to overlap of predictors and available treatment strategies.

Disclosure of interests

None declared. Completed disclosure of interests form available to view online as supporting information.

Contribution to authorship

The Expect Study I was designed by LS and MS. LM elaborated and carried out Expect Study I under the supervision of LS and HS. LM conducted the analyses, interpreted the data, and drafted the manuscript. LS and HS contributed to the interpretation of the outcomes and critically reviewed draft versions. SvK contributed to the imputation of the data and critically reviewed draft versions. RA, IvD, JL, IZ, and MS collaborated in data collection and critically reviewed draft versions. All authors gave approval of the final version of the manuscript.

Details of ethics approval

The Medical Ethics Committee (MEC) of the Maastricht University Medical Centre evaluated the study protocol and

declared that the study did not fall within the scope of the Dutch Medical Research Involving Human Subjects Act (WMO) (MEC 13-4-053). All participating women gave informed consent through the Internet. The study was registered at The Netherlands Trial Registry on 21 August 2013 (NTR4143, www.trialregister.nl).

Funding

The Expect Study I 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.

Acknowledgements

We thank all of the women who participated in the Expect Study I. The Expect Study I could not have been established without the contribution of the participating departments of obstetrics and gynaecology of hospitals, midwifery practices, and maternity care centres in the Province of Limburg: Zuyderland Medical Centre Heerlen and Sittard-Geleen, Maastricht University Medical Centre, Laurentius Hospital Roermond, Sint Jans Gasthuis Weert, VieCuri Medical Centre, 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 centre Echt.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Flowchart of validation cohort fetal growth deviations.

Figure S2. Receiver operating characteristic (ROC) curves of externally validated first-trimester prediction models for small-for-gestational-age infants with birthweights below the fifth and the tenth percentiles.

Figure S3. Calibration plots of recalibrated first-trimester prediction models for small-for-gestational-age infants with birthweights below the tenth percentile.

Figure S4. Receiver operating characteristic (ROC) curves of externally validated first-trimester prediction models for large-for-gestational-age infants with birthweights above the 90th and 95th percentiles.

Figure S5. Calibration plots of recalibrated first-trimester prediction models for large-for-gestational-age >90th percentile.

Table S1. Characteristics included prediction models for small-for-gestational-age.

Table S2. Characteristics included prediction models for large-for-gestational-age.

Table S3. Characteristics of observed and imputed validation cohort.

Appendix S1. Predictor assessment and model algorithms.

Appendix S2. Characteristics original cohorts and validation cohort. ■

References

1 Bjorstad AR, Irgens-Hansen K, Daltveit AK, Irgens LM. Macrosomia: mode of delivery and pregnancy outcome. *Acta Obstet Gynecol Scand* 2010;89:664–9.

2 Jolly MC, Sebire NJ, Harris JP, Regan L, Robinson S. Risk factors for macrosomia and its clinical consequences: a study of 350,311 pregnancies. *Eur J Obstet Gynecol Reprod Biol* 2003;111:9–14.

3 Weissmann-Brenner A, Simchen MJ, Zilberberg E, Kalter A, Weisz B, Achiron R, et al. Maternal and neonatal outcomes of large for gestational age pregnancies. *Acta Obstet Gynecol Scand* 2012;91:844–9.

4 Chavkin U, Wainstock T, Sheiner E, Sergienko R, Walfisch A. Perinatal outcome of pregnancies complicated with extreme birth weights at term. *J Matern Fetal Neonatal Med* 2019;32:198–202.

5 Mendez-Figueroa H, Truong VTT, Pedroza C, Chauhan SP. Large for gestational age infants and adverse outcomes among uncomplicated pregnancies at term. *Am J Perinatol* 2017;34:655–62.

6 Garite TJ, Clark R, Thorp JA. Intrauterine growth restriction increases morbidity and mortality among premature neonates. *Am J Obstet Gynecol* 2004;191:481–7.

7 Grisaru-Granovsky S, Reichman B, Lerner-Geva L, Boyko V, Hammerman C, Samueloff A, et al. Mortality and morbidity in preterm small-for-gestational-age infants: a population-based study. *Am J Obstet Gynecol* 2012;206:150 e1–7.

8 McIntire DD, Bloom SL, Casey BM, Leveno KJ. Birth weight in relation to morbidity and mortality among newborn infants. *N Engl J Med* 1999;340:1234–8.

9 Mendez-Figueroa H, Truong VT, Pedroza C, Chauhan SP. Morbidity and mortality in small-for-gestational-age infants: a secondary analysis of nine MFMU network studies. *Am J Perinatol* 2017;34:323–32.

10 Barker DJ, Osmond C, Forsen TJ, Kajantie E, Eriksson JG. Maternal and social origins of hypertension. *Hypertension* 2007;50:565–71.

11 Boney CM, Verma A, Tucker R, Vohr BR. Metabolic syndrome in childhood: association with birth weight, maternal obesity, and gestational diabetes mellitus. *Pediatrics* 2005;115:e290–6.

12 Crispi F, Miranda J, Gratacos E. Long-term cardiovascular consequences of fetal growth restriction: biology, clinical implications, and opportunities for prevention of adult disease. *Am J Obstet Gynecol* 2018;218:S869–79.

13 Eriksson JG, Forsen T, Tuomilehto J, Jaddoe VW, Osmond C, Barker DJ. Effects of size at birth and childhood growth on the insulin resistance syndrome in elderly individuals. *Diabetologia* 2002;45:342–8.

14 Harder T, Rodekamp E, Schellong K, Dudenhausen JW, Plagemann A. Birth weight and subsequent risk of type 2 diabetes: a meta-analysis. *Am J Epidemiol* 2007;165:849–57.

15 Hermann GM, Dallas LM, Haskell SE, Roghair RD. Neonatal macrosomia is an independent risk factor for adult metabolic syndrome. *Neonatology* 2010;98:238–44.

16 Newsome CA, Shiell AW, Fall CH, Phillips DI, Shier R, Law CM. Is birth weight related to later glucose and insulin metabolism?–A systematic review. *Diabet Med* 2003;20:339–48.

17 Ornoy A. Prenatal origin of obesity and their complications: gestational diabetes, maternal overweight and the paradoxical effects of fetal growth restriction and macrosomia. *Reprod Toxicol* 2011;32:205–12.

18 Rogers I, Group E-BS. The influence of birthweight and intrauterine environment on adiposity and fat distribution in later life. *Int J Obes Relat Metab Disord* 2003;27:755–77.

19 Johnston LB, Clark AJ, Savage MO. Genetic factors contributing to birth weight. *Arch Dis Child Fetal Neonatal Ed* 2002;86:F2–3.

20 Spencer N, Logan S. Social influences on birth weight. *Arch Dis Child Fetal Neonatal Ed* 2002;86:F6–7.

21 Stephenson T, Symonds ME. Maternal nutrition as a determinant of birth weight. *Arch Dis Child Fetal Neonatal Ed* 2002;86:F4–6.

22 He XJ, Qin FY, Hu CL, Zhu M, Tian CQ, Li L. Is gestational diabetes mellitus an independent risk factor for macrosomia: a meta-analysis? *Arch Gynecol Obstet* 2015;291:729–35.

23 Dai RX, He XJ, Hu CL. Maternal pre-pregnancy obesity and the risk of macrosomia: a meta-analysis. *Arch Gynecol Obstet* 2018;297:139–45.

24 Araujo Junior E, Peixoto AB, Zamarian AC, Elito Junior J, Tonni G. Macrosomia. *Best Pract Res Clin Obstet Gynaecol* 2017;38:83–96.

25 Anderson NH, Sadler LC, Stewart AW, Fyfe EM, McCowan LM. Independent risk factors for infants who are small for gestational age by customised birthweight centiles in a multi-ethnic New Zealand population. *Aust N Z J Obstet Gynaecol* 2013;53:136–42.

26 McCowan LM, Roberts CT, Dekker GA, Taylor RS, Chan EH, Kenny LC, et al. Risk factors for small-for-gestational-age infants by customised birthweight centiles: data from an international prospective cohort study. *BJOG* 2010;117:1599–607.

27 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.

28 Steyerberg EW, Moons KG, van dWD, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.

29 Boucoiran I, Djemli A, Taillefer C, Rypens F, Delvin E, Audibert F. First-trimester prediction of birth weight. *Am J Perinatol* 2013;30:665–72.

30 Crovetto F, Triunfo S, Crispi F, Rodriguez-Sureda V, Dominguez C, Figueras F, et al. Differential performance of first-trimester screening in predicting small-for-gestational-age neonate or fetal growth restriction. *Ultrasound Obstet Gynecol* 2017;49:349–56.

31 Frick AP, Syngelaki A, Zheng M, Poon LC, Nicolaides KH. Prediction of large-for-gestational-age neonates: screening by maternal factors and biomarkers in the three trimesters of pregnancy. *Ultrasound Obstet Gynecol* 2016;47:332–9.

32 Gonzalez Gonzalez NL, Plasencia W, Gonzalez Davila E, Padron E, di Renzo GC, Bartha JL. First and second trimester screening for large for gestational age infants. *J Matern Fetal Neonatal Med* 2013;26:1635–40.

33 Gonzalez-Gonzalez NL, Gonzalez-Davila E, Gonzalez Marrero L, Padron E, Conde JR, Plasencia W. Value of placental volume and vascular flow indices as predictors of intrauterine growth retardation. *Eur J Obstet Gynecol Reprod Biol* 2017;212:13–9.

34 Karagiannis G, Akolekar R, Sarquis R, Wright D, Nicolaides KH. Prediction of small-for-gestation neonates from biophysical and biochemical markers at 11-13 weeks. *Fetal Diagn Ther* 2011;29:148–54.

35 Leal AM, Poon LC, Frisova V, Veduta A, Nicolaides KH. First-trimester maternal serum tumor necrosis factor receptor-1 and pre-eclampsia. *Ultrasound Obstet Gynecol* 2009;33:135–41.

36 McCowan LM, Thompson JM, Taylor RS, Baker PN, North RA, Poston L, et al. Prediction of small for gestational age infants in healthy nulliparous women using clinical and ultrasound risk factors combined with early pregnancy biomarkers. *PLoS ONE* 2017;12:e0169311.

37 Nanda S, Akolekar R, Sarquis R, Mosconi AP, Nicolaides KH. Maternal serum adiponectin at 11 to 13 weeks of gestation in the prediction of macrosomia. *Prenat Diagn* 2011;31:479–83.

38 Onwudiwe N, Yu CK, Poon LC, Spiliopoulos I, Nicolaides KH. Prediction of pre-eclampsia by a combination of maternal history, uterine artery Doppler and mean arterial pressure. *Ultrasound Obstet Gynecol* 2008;32:877–83.

39 Papastefanou I, Souka AP, Pilalis A, Eleftheriades M, Michalitsi V, Kassanos D. First trimester prediction of small- and large-for-gestation neonates by an integrated model incorporating ultrasound parameters, biochemical indices and maternal characteristics. *Acta Obstet Gynecol Scand* 2012;91:104–11.

40 Plasencia W, Akolekar R, Dagklis T, Veduta A, Nicolaides KH. Placental volume at 11-13 weeks' gestation in the prediction of birth weight percentile. *Fetal Diagn Ther* 2011;30:23–8.

41 Plasencia W, Gonzalez Davila E, Tetilla V, Padron Perez E, Garcia Hernandez JA, Gonzalez Gonzalez NL. First-trimester screening for large-for-gestational-age infants. *Ultrasound Obstet Gynecol* 2012;39:389–95.

42 Poon LC, Karagiannis G, Stratieva V, Syngelaki A, Nicolaides KH. First-trimester prediction of macrosomia. *Fetal Diagn Ther* 2011;29:139–47.

43 Poon LC, Chelemen T, Granvillano O, Pandeva I, Nicolaides KH. First-trimester maternal serum a disintegrin and metalloprotease 12 (ADAM12) and adverse pregnancy outcome. *Obstet Gynecol* 2008;112:1082–90.

44 Poon LC, Karagiannis G, Staboulidou I, Shafiei A, Nicolaides KH. Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates. *Prenat Diagn* 2011;31:58–65.

45 Poon LC, Zaragoza E, Akolekar R, Anagnostopoulos E, Nicolaides KH. Maternal serum placental growth factor (PlGF) in small for gestational age pregnancy at 11(+0) to 13(+6) weeks of gestation. *Prenat Diagn* 2008;28:1110–5.

46 Schneuer FJ, Roberts CL, Ashton AW, Guilbert C, Tasevski V, Morris JM, et al. Angiopoietin 1 and 2 serum concentrations in first trimester of pregnancy as biomarkers of adverse pregnancy outcomes. *Am J Obstet Gynecol* 2014;210:345 e1–e9.

47 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.

48 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.

49 Visser GH, Eilers PH, Elferink-Stinkens PM, Merkus HM, Wit JM. New Dutch reference curves for birthweight by gestational age. *Early Hum Dev* 2009;85:737–44.

50 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.

51 Van Buuren S. *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC/Chapman & Hall; 2012.

52 Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer-Verlag; 2008.

53 Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;26:565–74.

54 Steyerberg EW, Vickers AJ. Decision curve analysis: a discussion. *Med Decis Making* 2008;28:146–9.

55 Schwartz N, Pessel C, Coletta J, Krieger AM, Timor-Tritsch IE. Early biometric lag in the prediction of small for gestational age neonates and preeclampsia. *J Ultrasound Med* 2011;30:55–60.

56 de Caunes F, Alexander GR, Berchel C, Guengant JP, Papiernik E. Anamnestic pregnancy risk assessment. *Int J Gynaecol Obstet* 1990;33:221–7.

57 Macdonald-Wallis C, Silverwood RJ, de Stavola BL, Inskip H, Cooper C, Godfrey KM, et al. Antenatal blood pressure for prediction of pre-eclampsia, preterm birth, and small for gestational age babies: development and validation in two general population cohorts. *BMJ* 2015;351:h5948.

58 Syngelaki A, Bredaki FE, Vaikousi E, Maiz N, Nicolaides KH. Body mass index at 11-13 weeks' gestation and pregnancy complications. *Fetal Diagn Ther* 2011;30:250–65.

59 Seed PT, Chappell LC, Black MA, Poppe KK, Hwang YC, Kasabov N, et al. Prediction of preeclampsia and delivery of small for gestational age babies based on a combination of clinical risk factors in high-risk women. *Hypertens Pregnancy* 2011;30:58–73.

60 Crovetto F, Crispi F, Scazzocchio E, Mercade I, Meler E, Figueras F, et al. First-trimester screening for early and late small-for-gestational-age neonates using maternal serum biochemistry, blood pressure and uterine artery Doppler. *Ultrasound Obstet Gynecol* 2014;43:34–40.

61 Poon LC, Syngelaki A, Akolekar R, Lai J, Nicolaides KH. Combined screening for preeclampsia and small for gestational age at 11-13 weeks. *Fetal Diagn Ther* 2013;33:16–27.

62 Berntorp K, Anderberg E, Claesson R, Ignell C, Kallen K. The relative importance of maternal body mass index and glucose levels for prediction of large-for-gestational-age births. *BMC Pregnancy Childbirth* 2015;15:280.

63 Mlynarczyk M, Chauhan SP, Baydoun HA, Wilkes CM, Earhart KR, Zhao Y, et al. The clinical significance of an estimated fetal weight below the 10th percentile: a comparison of outcomes of <5th vs 5th-9th percentile. *Am J Obstet Gynecol* 2017;217:198 e1–e11.

64 Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med Res Methodol* 2014;14:40.

65 Nardozza LM, Caetano AC, Zamarian AC, Mazzola JB, Silva CP, Marcal VM, et al. Fetal growth restriction: current knowledge. *Arch Gynecol Obstet* 2017;295:1061–77.

66 Mayer C, Joseph KS. Fetal growth: a review of terms, concepts and issues relevant to obstetrics. *Ultrasound Obstet Gynecol* 2013;41:136–45.

67 Al-Rubaie Z, Askie LM, Ray JG, Hudson HM, Lord SJ. The performance of risk prediction models for pre-eclampsia using routinely collected maternal characteristics and comparison with models that include specialised tests and with clinical guideline decision rules: a systematic review. *BJOG* 2016;123:1441–52.

68 Lamain – de Ruiter M, Kwee A, Naaktgeboren CA, Franx A, Moons KGM, Koster MPH. Prediction models for the risk of gestational diabetes: a systematic review. *Diagn Prognostic Res* 2017;1:3.

69 Meher S, Duley L, Hunter K, Askie L. Antiplatelet therapy before or after 16 weeks' gestation for preventing preeclampsia: an individual participant data meta-analysis. *Am J Obstet Gynecol* 2017;216:121–128 e2.

70 Roberge S, Nicolaides K, Demers S, Hyett J, Chaillet N, Bujold E. The role of aspirin dose on the prevention of preeclampsia and fetal growth restriction: systematic review and meta-analysis. *Am J Obstet Gynecol* 2017;216(110–20):e6.

71 Groom KM, David AL. The role of aspirin, heparin, and other interventions in the prevention and treatment of fetal growth restriction. *Am J Obstet Gynecol* 2018;218:S829–40.

72 Audette MC, Kingdom JC. Screening for fetal growth restriction and placental insufficiency. *Semin Fetal Neonatal Med* 2018;23:119–25.

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

74 Bond DM, Gordon A, Hyett J, de VB, Carberry AE, Morris J. Planned early delivery versus expectant management of the term suspected compromised baby for improving outcomes. *Cochrane Database Syst Rev* 2015;11:CD009433.

75 Gardosi J, Madurasinghe V, Williams M, Malik A, Francis A. Maternal and fetal risk factors for stillbirth: population based study. *BMJ* 2013;346:f108.

76 Boulvain M, Irion O, Dowswell T, Thornton JG. Induction of labour at or near term for suspected fetal macrosomia. *Cochrane Database Syst Rev* 2016;5:CD000938.

77 Magro-Malosso ER, Saccone G, Chen M, Navathe R, Di Tommaso M, Berghella V. Induction of labour for suspected macrosomia at term in non-diabetic women: a systematic review and meta-analysis of randomized controlled trials. *BJOG* 2017;124:414–21.

78 Plasencia W, Maiz N, Bonino S, Kaihura C, Nicolaides KH. Uterine artery Doppler at 11 + 0 to 13 + 6 weeks in the prediction of pre-eclampsia. *Ultrasound Obstet Gynecol* 2007;30:742–9.