NIHR | National Institute for Health and Care Research

Journals Library

Check for updates

Health Technology Assessment

Volume 28 • Issue 47 • August 2024

ISSN 2046-4924

Development and validation of prediction models for fetal growth restriction and birthweight: an individual participant data meta-analysis

John Allotey, Lucinda Archer, Dyuti Coomar, Kym IE Snell, Melanie Smuk, Lucy Oakey, Sadia Haqnawaz, Ana Pilar Betrán, Lucy C Chappell, Wessel Ganzevoort, Sanne Gordijn, Asma Khalil, Ben W Mol, Rachel K Morris, Jenny Myers, Aris T Papageorghiou, Basky Thilaganathan, Fabricio Da Silva Costa, Fabio Facchinetti, Arri Coomarasamy, Akihide Ohkuchi, Anne Eskild, Javier Arenas Ramírez, Alberto Galindo, Ignacio Herraiz, Federico Prefumo, Shigeru Saito, Line Sletner, Jose Guilherme Cecatti, Rinat Gabbay-Benziv, Francois Goffinet, Ahmet A Baschat, Renato T Souza, Fionnuala Mone, Diane Farrar, Seppo Heinonen, Kjell Å Salvesen, Luc JM Smits, Sohinee Bhattacharya, Chie Nagata, Satoru Takeda, Marleen MHJ van Gelder, Dewi Anggraini, SeonAe Yeo, Jane West, Javier Zamora, Hema Mistry, Richard D Riley and Shakila Thangaratinam for the IPPIC Collaborative Network

DOI 10.3310/DABW4814

Development and validation of prediction models for fetal growth restriction and birthweight: an individual participant data meta-analysis

John Allotey,[1*] Lucinda Archer,[2] Dyuti Coomar,[1] Kym IE Snell,[2] Melanie Smuk,[3] Lucy Oakey,[1] Sadia Haqnawaz,[4] Ana Pilar Betrán,[5] Lucy C Chappell,[6] Wessel Ganzevoort,[7] Sanne Gordijn,[8] Asma Khalil,[9] Ben W Mol,[10,11] Rachel K Morris,[12] Jenny Myers,[13] Aris T Papageorghiou,[9] Basky Thilaganathan,[9,14] Fabricio Da Silva Costa,[15] Fabio Facchinetti,[16] Arri Coomarasamy,[1] Akihide Ohkuchi,[17] Anne Eskild,[18] Javier Arenas Ramírez,[19] Alberto Galindo,[20] Ignacio Herraiz,[21] Federico Prefumo,[22] Shigeru Saito,[23] Line Sletner,[24] Jose Guilherme Cecatti,[25] Rinat Gabbay-Benziv,[26] Francois Goffinet,[27,28] Ahmet A Baschat,[29] Renato T Souza,[25] Fionnuala Mone,[30] Diane Farrar,[31] Seppo Heinonen,[32] Kjell Å Salvesen,[33] Luc JM Smits,[34] Sohinee Bhattacharya,[11] Chie Nagata,[35] Satoru Takeda,[36] Marleen MHJ van Gelder,[37] Dewi Anggraini,[38] SeonAe Yeo,[39] Jane West,[31] Javier Zamora,[1,40] Hema Mistry,[41] Richard D Riley[2] and Shakila Thangaratinam[1,42] for the IPPIC Collaborative Network

[1]WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK
[2]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK
[3]Blizard Institute, Centre for Genomics and Child Health, Queen Mary University of London, London, UK
[4]The Hildas, Dame Hilda Lloyd Network, WHO Collaborating Centre for Global Women's Health, University of Birmingham, Birmingham, UK
[5]Department of Reproductive and Health Research, World Health Organization, Geneva, Switzerland

[6]Department of Women and Children's Health, School of Life Course Sciences, King's College London, London, UK
[7]Department of Obstetrics, Amsterdam UMC University of Amsterdam, Amsterdam, the Netherlands
[8]Faculty of Medical Sciences, University Medical Center Groningen, Groningen, the Netherlands
[9]Fetal Medicine Unit, St George's University Hospitals NHS Foundation Trust and Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK
[10]Department of Obstetrics and Gynaecology, Monash University, Monash Medical Centre, Clayton, Victoria, Australia
[11]Aberdeen Centre for Women's Health Research, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
[12]Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[13]Maternal and Fetal Health Research Centre, Manchester Academic Health Science Centre, University of Manchester, Central Manchester NHS Trust, Manchester, UK
[14]Tommy's National Centre for Maternity Improvement, Royal College of Obstetrics and Gynaecology, London, UK
[15]Maternal Fetal Medicine Unit, Gold Coast University Hospital and School of Medicine, Griffith University, Gold Coast, Queensland, Australia
[16]Mother-Infant Department, University of Modena and Reggio Emilia, Emilia-Romagna, Italy
[17]Department of Obstetrics and Gynecology, Jichi Medical University School of Medicine, Shimotsuke-shi, Tochigi, Japan
[18]Akershus University Hospital, University of Oslo, Oslo, Norway
[19]Hospital Universitario de Cabueñes, Gijón, Spain
[20]Fetal Medicine Unit, Maternal and Child Health and Development Network (SAMID), Department of Obstetrics and Gynaecology, Hospital Universitario, Instituto de Investigación Hospital, Universidad Complutense de Madrid, Madrid, Spain
[21]Department of Obstetrics and Gynaecology, Hospital Universitario, Madrid, Spain
[22]Department of Clinical and Experimental Sciences, University of Brescia, Italy
[23]Department Obstetrics and Gynecology, University of Toyama, Toyama, Japan
[24]Deptartment of Pediatric and Adolescents Medicine, Akershus University Hospital, Sykehusveien, Norway
[25]Obstetric Unit, Department of Obstetrics and Gynecology, University of Campinas, Campinas, Sao Paulo, Brazil
[26]Maternal Fetal Medicine Unit, Department of Obstetrics and Gynecology, Hillel Yaffe Medical Center Hadera, Affiliated to the Ruth and Bruce Rappaport School of Medicine, Technion, Haifa, Israel
[27]Maternité Port-Royal, AP-HP, APHP, Centre-Université de Paris, FHU PREMA, Paris, France
[28]Université de Paris, INSERM U1153, Equipe de recherche en Epidémiologie Obstétricale, Périnatale et Pédiatrique (EPOPé), Centre de Recherche Epidémiologie et Biostatistique Sorbonne Paris Cité (CRESS), Paris, France
[29]Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, MD, USA
[30]Centre for Public Health, Queen's University, Belfast, UK
[31]Bradford Institute for Health Research, Bradford, UK

[32]Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[33]Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway
[34]Care and Public Health Research Institute, Maastricht University Medical Centre, Maastricht, the Netherlands
[35]Center for Postgraduate Education and Training, National Center for Child Health and Development, Tokyo, Japan
[36]Department of Obstetrics and Gynecology, Juntendo University, Tokyo, Japan
[37]Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands
[38]Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, South Kalimantan, Indonesia
[39]University of North Carolina at Chapel Hill, School of Nursing, NC, USA
[40]Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal (IRYCIS), Madrid, Spain
[41]Warwick Medical School, University of Warwick, Warwick, UK
[42]Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK

*Corresponding author

Disclosure of interests

Full disclosure of interests: Completed ICMJE forms for all authors, including all related interests, are available in the toolkit on the NIHR Journals Library report publication page at https://doi.org/10.3310/DABW4814.

Primary conflicts of interest: Basky Thilaganathan reports funding from Tommy's Fund award and iPLACENTA. Fabricio Da Silva Costa is a member of the International Society of Ultrasound in Obstetrics and Gynecology (ISUOG). Lucy C Chappell is Chief Executive Officer of the National Institute for Health and Care Research (NIHR). Asma Khalil is a member of the NIHR Funding Committee, an ISUOG trustee and Obstetric Lead at the National Maternity and Perinatal Audit (NMPA). Rachel K Morris is lead developer for Royal College of Obstetricians and Gynaecologists (RCOG) guideline on fetal growth restriction and a member of the Saving Babies Care Bundle steering group. Richard D Riley reports payment for Statistical Reviews for the *BMJ* and occasionally other journals, guest lecturer at McGill and royalties for textbooks edited. Ben W Mol is supported by a NHMRC Investigator grant (GNT1176437), reports consultancy for ObsEva and has received research funding from Ferring and Merck. Aris T Papageorghiou is supported by the NIHR Oxford Biomedical Research Centre (BRC). SeonAe Yeo reports grants from National Institute of Nursing Research (NINR). Alberto Galindo reports grants from Instituto de Salud Carlos III (Spanish Ministry of Economy, Industry and Competitiveness) and payment from Roche Diagnostics for lectures and expert advisory board membership. Ignacio Herraiz reports grants from Instituto de Salud Carlos III (Spanish Ministry of Economy, Industry and Competitiveness) and payment from Roche Diagnostics and Thermo-Fischer for lectures and expert advisory board membership. Sanne Gordijn reports payment from Roche as in-kind supply of sFLt/PLGF for fetal growth restriction studies CEPRA and grant from ZonMw. Arri Coomarasamy is a Funding Committee Member for EME.

Other authors do not report any competing interests.

Published August 2024
DOI: 10.3310/DABW4814

This report should be referenced as follows:

Allotey J, Archer L, Coomar D, Snell KIE, Smuk M, Oakey L, *et al*. Development and validation of prediction models for fetal growth restriction and birthweight: an individual participant data meta-analysis. *Health Technol Assess* 2024;**28**(47). https://doi.org/10.3310/DABW4814

Health Technology Assessment

ISSN 2046-4924 (Online)

Impact factor: 3.6

A list of Journals Library editors can be found on the NIHR Journals Library website

Launched in 1997, *Health Technology Assessment* (HTA) has an impact factor of 3.6 and is ranked 32nd (out of 105 titles) in the 'Health Care Sciences & Services' category of the Clarivate 2022 Journal Citation Reports (Science Edition). It is also indexed by MEDLINE, CINAHL (EBSCO Information Services, Ipswich, MA, USA), EMBASE (Elsevier, Amsterdam, the Netherlands), NCBI Bookshelf, DOAJ, Europe PMC, the Cochrane Library (John Wiley & Sons, Inc., Hoboken, NJ, USA), INAHTA, the British Nursing Index (ProQuest LLC, Ann Arbor, MI, USA), Ulrichsweb™ (ProQuest LLC, Ann Arbor, MI, USA) and the Science Citation Index Expanded™ (Clarivate™, Philadelphia, PA, USA).

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta.

Criteria for inclusion in the *Health Technology Assessment* journal
Manuscripts are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

Health Technology Assessment (HTA) research is undertaken where some evidence already exists to show that a technology can be effective and this needs to be compared to the current standard intervention to see which works best. Research can evaluate any intervention used in the treatment, prevention or diagnosis of disease, provided the study outcomes lead to findings that have the potential to be of direct benefit to NHS patients. Technologies in this context mean any method used to promote health; prevent and treat disease; and improve rehabilitation or long-term care. They are not confined to new drugs and include any intervention used in the treatment, prevention or diagnosis of disease.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

This article

The research reported in this issue of the journal was funded by the HTA programme as award number 17/148/07. The contractual start date was in May 2019. The draft manuscript began editorial review in January 2022 and was accepted for publication in September 2022. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' manuscript and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this article.

This article presents independent research funded by the National Institute for Health and Care Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care.

This article was published based on current knowledge at the time and date of publication. NIHR is committed to being inclusive and will continually monitor best practice and guidance in relation to terminology and language to ensure that we remain relevant to our stakeholders.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Newgen Digitalworks Pvt Ltd, Chennai, India (www.newgen.co).

Abstract

Development and validation of prediction models for fetal growth restriction and birthweight: an individual participant data meta-analysis

John Allotey,[1*] Lucinda Archer,[2] Dyuti Coomar,[1] Kym IE Snell,[2] Melanie Smuk,[3] Lucy Oakey,[1] Sadia Haqnawaz,[4] Ana Pilar Betrán,[5] Lucy C Chappell,[6] Wessel Ganzevoort,[7] Sanne Gordijn,[8] Asma Khalil,[9] Ben W Mol,[10,11] Rachel K Morris,[12] Jenny Myers,[13] Aris T Papageorghiou,[9] Basky Thilaganathan,[9,14] Fabricio Da Silva Costa,[15] Fabio Facchinetti,[16] Arri Coomarasamy,[1] Akihide Ohkuchi,[17] Anne Eskild,[18] Javier Arenas Ramírez,[19] Alberto Galindo,[20] Ignacio Herraiz,[21] Federico Prefumo,[22] Shigeru Saito,[23] Line Sletner,[24] Jose Guilherme Cecatti,[25] Rinat Gabbay-Benziv,[26] Francois Goffinet,[27,28] Ahmet A Baschat,[29] Renato T Souza,[25] Fionnuala Mone,[30] Diane Farrar,[31] Seppo Heinonen,[32] Kjell Å Salvesen,[33] Luc JM Smits,[34] Sohinee Bhattacharya,[11] Chie Nagata,[35] Satoru Takeda,[36] Marleen MHJ van Gelder,[37] Dewi Anggraini,[38] SeonAe Yeo,[39] Jane West,[31] Javier Zamora,[1,40] Hema Mistry,[41] Richard D Riley[2] and Shakila Thangaratinam[1,42] for the IPPIC Collaborative Network

[1]WHO Collaborating Centre for Global Women's Health, Institute of Metabolism and Systems Research, University of Birmingham, Birmingham, UK
[2]Centre for Prognosis Research, School of Medicine, Keele University, Keele, UK
[3]Blizard Institute, Centre for Genomics and Child Health, Queen Mary University of London, London, UK
[4]The Hildas, Dame Hilda Lloyd Network, WHO Collaborating Centre for Global Women's Health, University of Birmingham, Birmingham, UK
[5]Department of Reproductive and Health Research, World Health Organization, Geneva, Switzerland
[6]Department of Women and Children's Health, School of Life Course Sciences, King's College London, London, UK
[7]Department of Obstetrics, Amsterdam UMC University of Amsterdam, Amsterdam, the Netherlands
[8]Faculty of Medical Sciences, University Medical Center Groningen, Groningen, the Netherlands
[9]Fetal Medicine Unit, St George's University Hospitals NHS Foundation Trust and Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK
[10]Department of Obstetrics and Gynaecology, Monash University, Monash Medical Centre, Clayton, Victoria, Australia
[11]Aberdeen Centre for Women's Health Research, Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, UK
[12]Institute of Applied Health Research, University of Birmingham, Birmingham, UK
[13]Maternal and Fetal Health Research Centre, Manchester Academic Health Science Centre, University of Manchester, Central Manchester NHS Trust, Manchester, UK
[14]Tommy's National Centre for Maternity Improvement, Royal College of Obstetrics and Gynaecology, London, UK
[15]Maternal Fetal Medicine Unit, Gold Coast University Hospital and School of Medicine, Griffith University, Gold Coast, Queensland, Australia
[16]Mother-Infant Department, University of Modena and Reggio Emilia, Emilia-Romagna, Italy

[17]Department of Obstetrics and Gynecology, Jichi Medical University School of Medicine, Shimotsuke-shi, Tochigi, Japan
[18]Akershus University Hospital, University of Oslo, Oslo, Norway
[19]Hospital Universitario de Cabueñes, Gijón, Spain
[20]Fetal Medicine Unit, Maternal and Child Health and Development Network (SAMID), Department of Obstetrics and Gynaecology, Hospital Universitario, Instituto de Investigación Hospital, Universidad Complutense de Madrid, Madrid, Spain
[21]Department of Obstetrics and Gynaecology, Hospital Universitario, Madrid, Spain
[22]Department of Clinical and Experimental Sciences, University of Brescia, Italy
[23]Department Obstetrics and Gynecology, University of Toyama, Toyama, Japan
[24]Deptartment of Pediatric and Adolescents Medicine, Akershus University Hospital, Sykehusveien, Norway
[25]Obstetric Unit, Department of Obstetrics and Gynecology, University of Campinas, Campinas, Sao Paulo, Brazil
[26]Maternal Fetal Medicine Unit, Department of Obstetrics and Gynecology, Hillel Yaffe Medical Center Hadera, Affiliated to the Ruth and Bruce Rappaport School of Medicine, Technion, Haifa, Israel
[27]Maternité Port-Royal, AP-HP, APHP, Centre-Université de Paris, FHU PREMA, Paris, France
[28]Université de Paris, INSERM U1153, Equipe de recherche en Epidémiologie Obstétricale, Périnatale et Pédiatrique (EPOPé), Centre de Recherche Epidémiologie et Biostatistique Sorbonne Paris Cité (CRESS), Paris, France
[29]Department of Gynecology and Obstetrics, Johns Hopkins University School of Medicine, MD, USA
[30]Centre for Public Health, Queen's University, Belfast, UK
[31]Bradford Institute for Health Research, Bradford, UK
[32]Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[33]Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway
[34]Care and Public Health Research Institute, Maastricht University Medical Centre, Maastricht, the Netherlands
[35]Center for Postgraduate Education and Training, National Center for Child Health and Development, Tokyo, Japan
[36]Department of Obstetrics and Gynecology, Juntendo University, Tokyo, Japan
[37]Department for Health Evidence, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands
[38]Faculty of Mathematics and Natural Sciences, Lambung Mangkurat University, South Kalimantan, Indonesia
[39]University of North Carolina at Chapel Hill, School of Nursing, NC, USA
[40]Clinical Biostatistics Unit, Hospital Universitario Ramón y Cajal (IRYCIS), Madrid, Spain
[41]Warwick Medical School, University of Warwick, Warwick, UK
[42]Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK

*Corresponding author j.allotey.1@bham.ac.uk

Background: Fetal growth restriction is associated with perinatal morbidity and mortality. Early identification of women having at-risk fetuses can reduce perinatal adverse outcomes.

Objectives: To assess the predictive performance of existing models predicting fetal growth restriction and birthweight, and if needed, to develop and validate new multivariable models using individual participant data.

Design: Individual participant data meta-analyses of cohorts in International Prediction of Pregnancy Complications network, decision curve analysis and health economics analysis.

Participants: Pregnant women at booking.

External validation of existing models (9 cohorts, 441,415 pregnancies); International Prediction of Pregnancy Complications model development and validation (4 cohorts, 237,228 pregnancies).

Predictors: Maternal clinical characteristics, biochemical and ultrasound markers.

Primary outcomes:

1. fetal growth restriction defined as birthweight <10th centile adjusted for gestational age and with stillbirth, neonatal death or delivery before 32 weeks' gestation
2. birthweight.

Analysis: First, we externally validated existing models using individual participant data meta-analysis. If needed, we developed and validated new International Prediction of Pregnancy Complications models using random-intercept regression models with backward elimination for variable selection and undertook internal-external cross-validation. We estimated the study-specific performance (*c*-statistic, calibration slope, calibration-in-the-large) for each model and pooled using random-effects meta-analysis. Heterogeneity was quantified using τ^2 and 95% prediction intervals. We assessed the clinical utility of the fetal growth restriction model using decision curve analysis, and health economics analysis based on National Institute for Health and Care Excellence 2008 model.

Results: Of the 119 published models, one birthweight model (Poon) could be validated. None reported fetal growth restriction using our definition. Across all cohorts, the Poon model had good summary calibration slope of 0.93 (95% confidence interval 0.90 to 0.96) with slight overfitting, and underpredicted birthweight by 90.4 g on average (95% confidence interval 37.9 g to 142.9 g).

The newly developed International Prediction of Pregnancy Complications-fetal growth restriction model included maternal age, height, parity, smoking status, ethnicity, and any history of hypertension, pre-eclampsia, previous stillbirth or small for gestational age baby and gestational age at delivery. This allowed predictions conditional on a range of assumed gestational ages at delivery. The pooled apparent *c*-statistic and calibration were 0.96 (95% confidence interval 0.51 to 1.0), and 0.95 (95% confidence interval 0.67 to 1.23), respectively. The model showed positive net benefit for predicted probability thresholds between 1% and 90%.

In addition to the predictors in the International Prediction of Pregnancy Complications-fetal growth restriction model, the International Prediction of Pregnancy Complications-birthweight model included maternal weight, history of diabetes and mode of conception. Average calibration slope across cohorts in the internal-external cross-validation was 1.00 (95% confidence interval 0.78 to 1.23) with no evidence of overfitting. Birthweight was underestimated by 9.7 g on average (95% confidence interval −154.3 g to 173.8 g).

Limitations: We could not externally validate most of the published models due to variations in the definitions of outcomes. Internal-external cross-validation of our International Prediction of Pregnancy Complications-fetal growth restriction model was limited by the paucity of events in the included cohorts. The economic evaluation using the published National Institute for Health and Care Excellence 2008 model may not reflect current practice, and full economic evaluation was not possible due to paucity of data.

Future work: International Prediction of Pregnancy Complications models' performance needs to be assessed in routine practice, and their impact on decision-making and clinical outcomes needs evaluation.

Conclusion: The International Prediction of Pregnancy Complications-fetal growth restriction and International Prediction of Pregnancy Complications-birthweight models accurately predict fetal growth restriction and birthweight for various assumed gestational ages at delivery. These can be used to stratify the risk status at booking, plan monitoring and management.

Study registration: This study is registered as PROSPERO CRD42019135045.

Funding: This award was funded by the National Institute for Health and Care Research (NIHR) Health Technology Assessment programme (NIHR award ref: 17/148/07) and is published in full in *Health Technology Assessment*; Vol. 28, No. 47. See the NIHR Funding and Awards website for further award information.

Contents

List of tables

List of figures

List of abbreviations

AC	abdominal circumference
ACOG	American College of Obstetricians and Gynecologists
ALSPAC	Avon Longitudinal Study of Parents and Children
CENTRAL	Cochrane Central Register of Controlled Trials
CI	confidence intervals
CITL	calibration-in-the-large
CS	caesarean section
DCA	decision curve analysis
EFW	estimated fetal weight
FGR	fetal growth restriction
FN	false negative
FP	false positive
GA	gestational age at delivery
GROW	gestation-related optimal weight
HRQoL	health-related quality of life
IECV	internal-external cross-validation
IPD	individual participant data
IPPIC	International Prediction of Pregnancy Complications
IQR	interquartile range
IUGR	intrauterine growth restriction
JSOG	Japan Society of Obstetrics and Gynecology
LP	linear predictor
MAR	missing at random
MFP	multivariable fractional polynomial
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NICHD CSL	National Institute of Child Health and Human Development Consortium on Safe Labour
PE	pre-eclampsia
PROBAST	Prediction study Risk of Bias Assessment Tool
PSS	personal social services
QALY	quality-adjusted life-year
RCOG	Royal College of Obstetrics and Gynaecology
SD	standard deviation
SFH	symphysis-fundal height
SGA	small for gestational age
TN	true negative
TP	true positive

Plain language summary

One in ten babies is born small for their age. A third of such small babies are considered to be 'growth-restricted' as they have complications such as dying in the womb (stillbirth) or after birth (newborn death), cerebral palsy, or needing long stays in hospital. When growth restriction is suspected in fetuses, they are closely monitored and often delivered early to avoid complications. Hence, it is important that we identify growth-restricted babies early to plan care.

Our goal was to provide personalised and accurate estimates of the mother's chances of having a growth-restricted baby and predict the baby's weight if delivered at various time points in pregnancy. To do so, first we tested how accurate existing risk calculators ('prediction models') were in predicting growth restriction and birthweight. We then developed new risk-calculators and studied their clinical and economic benefits. We did so by accessing the data from individual pregnant women and their babies in our large database library (International Prediction of Pregnancy Complications).

Published risk-calculators had various definitions of growth restriction and none predicted the chances of having a growth-restricted baby using our definition. One predicted baby's birthweight. This risk-calculator performed well, but underpredicted the birthweight by up to 143 g.

We developed two new risk-calculators to predict growth-restricted babies (International Prediction of Pregnancy Complications-fetal growth restriction) and birthweight (International Prediction of Pregnancy Complications-birthweight). Both calculators accurately predicted the chances of the baby being born with growth restriction, and its birthweight. The birthweight was underpredicted by <9.7 g. The calculators performed well in both mothers predicted to be low and high risk.

Further research is needed to determine the impact of using these calculators in practice, and challenges to implementing them in practice. Both International Prediction of Pregnancy Complications-fetal growth restriction and International Prediction of Pregnancy Complications-birthweight risk calculators will inform healthcare professionals and empower parents make informed decisions on monitoring and timing of delivery.

Scientific summary

Background

Fetal growth restriction (FGR) is associated with perinatal mortality and morbidity. Early and accurate identification and appropriate management of pregnant women with growth-restricted fetuses can reduce perinatal complications.

Objectives

Primary

Using individual personal data (IPD) meta-analysis

1. To externally validate the predictive accuracy of existing prediction models for FGR (birthweight < 10th centile adjusted for gestational age, with serious perinatal complications such as stillbirth, neonatal death or delivery before 32 weeks), and birthweight within cohorts in the International Prediction of Pregnancy Complications (IPPIC) data repository.
2. To develop and validate [using internal-external cross-validation (IECV)] new multivariable prediction models for (1) FGR and (2) birthweight at various potential gestational ages of delivery.

Secondary

1. To compare the predictive performance of models according to (1) population (selected – high/low risk; unselected); (2) trimester of testing (first <14 weeks; second ~20 weeks; third ~28 weeks); (3) choice of predictors (clinical only; clinical and ultrasound; clinical and biochemical; clinical, ultrasound and biochemical); and (4) onset of FGR (early <32 weeks; late >32 weeks).
2. To assess if the performance of the prediction models is generalisable for various definitions of FGR, and assess the association between various birthweight centiles calculated using customised and population-based standards and perinatal morbidity and mortality.
3. To estimate the net benefit (clinical utility) of the developed prediction models using decision curve analysis (DCA).
4. To assess the costs and outcomes and the potential impact of resource use of the prediction models.

Methods

We followed existing recommendations for prediction model development and validation and reported in line with guidelines for prognostic research and IPD meta-analysis.

Our meta-analysis utilised IPD within the IPPIC Network database. IPPIC is a living data repository of cleaned and harmonised data of pregnant women from large birth or population-based cohorts, study cohort data, registries or unpublished data from hospital records. The primary outcomes were (1) FGR defined as birthweight < 10th centile adjusted for gestational age, with serious complications such as stillbirth, neonatal death, or delivery before 32 weeks and (2) birthweight for deliveries at various potential gestational ages.

We updated our previous searches (inception to July 2012) for relevant prediction models published until August 2019 for external validation. Models were validated if at least one IPPIC IPD cohort contained all the predictors included in the model, and the model outcome occurred in some of the participants in the IPD cohort. Partially missing predictors and outcome variables missing for < 90% of

individuals in the cohorts were imputed using multiple imputation by chained equations, assuming that individual values were missing at random. Imputation was performed separately for each cohort to allow for the clustering of individuals within cohorts. The predictive performance of existing model was evaluated using measures of calibration (agreement between predicted and observed outcomes), and discrimination (how well model differentiates between those with and without the outcome, ideal value 1) for each cohort separately and then pooled using a random-effects model estimated using restricted maximum likelihood.

Candidate predictors for development of FGR and birthweight models were identified following a prioritisation survey by clinical experts and from existing prediction models. Prediction models were developed using random intercept regression models with backward elimination for variable selection, and IECV was used for validation. Model predictive performance measures [calibration-in-the-large (CITL), the calibration slope, the *c*-statistic and Nagelkerke's R^2] were summarised using random-effects meta-analysis to give a pooled estimate of overall performance across cohorts.

We assessed the clinical utility of IPPIC-FGR model using DCA. By weighing up potential benefit and harm, the net benefit of the model was plotted at various clinically relevant threshold probabilities. Decision curves were compared against 'treat-all' and 'treat-none' strategies across the range of predicted threshold probabilities at which the model may be clinically useful. We also evaluated the costs and outcomes of IPPIC-FGR model using a decision analytical model constructed using Microsoft Excel®. The costs and outcomes of IPPIC-FGR model was compared against existing strategies in the National Institute for Health and Care Excellence (NICE) 2008 Antenatal Care guideline [no monitoring for FGR and monitoring FGR of all fetuses using ultrasound and symphysis-fundal height (SFH) measurement]. Costs were from the perspective of the National Health Service, and no discounting was required due to the short timeframe from entry into the model to outcome.

Results

External validation of existing prediction models

Overall, 119 published prediction models (55 articles) for FGR and birthweight were identified, with various definitions of FGR or birthweight outcome dichotomised. No study reported our predefined outcome of FGR. Of the eleven models that predicted birthweight on a continuous scale, only one (Poon 2011; 33,602 pregnancies) reported variables available in the IPPIC cohorts and was externally validated in nine IPPIC cohorts involving 441,415 pregnancies. The Poon model included gestational age at delivery, maternal weight, height, age, parity, smoking status, ethnicity, history of chronic hypertension, diabetes and assisted conception. Calibration slopes of the model ranged from 0.91 to 1.05, with a pooled calibration slope across all cohorts of 0.974 [95% confidence interval (CI) 0.938 to 1.011, $\tau^2 = 0.0018$]. On average, the model systematically underpredicted birthweight by 90.4 g (37.9 g to 142.9 g) across the validation cohorts and showed moderate heterogeneity in performance.

Development and validation of IPPIC-FGR and IPPIC-birthweight models

We developed the IPPIC-FGR model using data from four IPPIC cohorts (237,228 pregnancies). The model included gestational age at delivery, mother's age, mother's height, parity, smoking status, ethnicity, history of hypertension, and any history of pre-eclampsia, stillbirth or small for gestational age baby. The pooled apparent *c*-statistic was 0.96 (95% CI 0.51 to 1.0), and the pooled apparent calibration slope was 0.95 (95% CI 0.67 to 1.23).

The IPPIC-birthweight model additionally included maternal weight, a history of diabetes and mode of conception, and was developed in same four IPPIC cohorts as for the IPPIC-FGR model. The pooled calibration slope across cohorts in the IECV was 1.0 (95% CI 0.78 to 1.23), thus showing no evidence of overfitting. Underestimation of birthweight was by 9.7 g on average across cohorts in the IECV (95% CI −154.3 g to 173.8 g) as assessed by CITL.

Decision curve analysis

The IPPIC-FGR model showed positive net benefit for predicted probability thresholds between 1% and 90% across all cohorts compared to a strategy of managing all pregnant women as if they will have growth-restricted fetuses, or managing them as if none will have growth-restricted fetuses (i.e. treat-all or treat-none strategies). Net benefit was greatest when the model was used in pregnancies <32 weeks' gestation. While there was no overall benefit in using the IPPIC-FGR model in pregnancies at or above 32 weeks' gestation compared to a strategy of treat-all, use of the model in pregnant women at this gestational age resulted in no additional harm in these group of women.

Health economics analysis

The health economics analysis based on NICE 2008 economic model for monitoring fetal growth showed the use of the IPPIC-FGR model was slightly more costly, and more perinatal deaths were saved for every 1000 FGR babies than the alternate strategy of no screening for FGR. When the IPPIC-FGR model was compared with screening using only SFH and ultrasound, the strategy was cheaper and again more perinatal deaths were prevented. Sensitivity analysis found that the results were robust and in line with the base-case analyses. The economic model did not take into account current pathways used to screen women at high risk of having FGR babies.

Recommendations for clinical practice and research

Incorporation of personalised predicted birthweight estimates (for various potential gestational ages) within existing growth charts, and risk stratification at booking for FGR can help plan intensity of fetal monitoring and timing of delivery. The impact of using IPPIC-FGR and IPPIC-birthweight models on changes in clinical practice and clinical outcomes needs further evaluation. Qualitative data are needed to determine the barriers and facilitators of their routine implementation in clinical practice. Our health economics analysis was based on the 2008 NICE model which is no longer reflective of current management strategies for risk assessing FGR. Therefore, in light of significant changes to current guidelines and care pregnant women at risk of FGRs receive, a detailed full economic evaluation is needed, which evaluates various strategies to risk assess FGR along current care pathways.

Conclusion

IPPIC-FGR and IPPIC-birthweight models accurately predict FGR and birthweight. The latter has better calibration than existing model. IPPIC-FGR model use is cost-effective. Both IPPIC models can help plan intensity of fetal monitoring in pregnancy and timing of delivery, to minimise adverse perinatal outcomes.

Study registration

This study is registered as PROSPERO CRD42019135045.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Health Technology Assessment programme (NIHR award ref: 17/148/07) and is published in full in *Health Technology Assessment*; Vol. 28, No. 47. See the NIHR Funding and Awards website for further award information.

Chapter 1 Background

Fetal growth restriction (FGR) or intrauterine growth restriction (IUGR) is defined as the failure of a fetus to achieve its intrinsically determined growth potential.[1] It is associated with perinatal morbidity and mortality, and long-term offspring complications such as neurodevelopmental delay, poor growth, adult-onset diseases in infancy and adolescence, including obesity, metabolic syndrome, type 2 diabetes and cardiovascular diseases.[2–4]

Fetal growth restriction is often used interchangeably with 'small for gestational age' (SGA),[5] where the estimated fetal weight (EFW) or birthweight of the fetus is <10th centile. However, of the 70,000 babies born small each year in England and Wales, up to 70% are constitutionally small, without major complications.[6] But one in three small babies is growth restricted, with arrest or shift in rates of growth trajectory, which increases their risk of immediate and long-term complications.[7,8] The odds of stillbirth (OR 7.1–10.0) and neonatal death (OR 3.4–9.4) are significantly higher in growth-restricted compared to normal weight fetuses at every week beyond the expected date of delivery in these babies.[2] Of the 3000 babies who were stillborn each year in the United Kingdom (UK), half were considered to be growth restricted.[8]

In growth-restricted fetuses, the condition is diagnosed early (<32 weeks) and is usually associated with hypertensive disorders of pregnancy and severe placental pathology.[9] These infants are often delivered early, with additional prematurity-related complications. Many cases of FGR are of late-onset (>32 weeks). The diagnosis is missed in three-quarters of these babies.[10] Early identification of women at risk of FGR can reduce perinatal mortality and morbidity, by identifying women who need close monitoring in pregnancy, and to plan the setting and timing of delivery to minimise adverse perinatal outcomes.

Considerable variation exists between international guidelines on how identify women at risk of having FGR. This ranges from arbitrarily chosen 'major' or 'minor' clinical risk factors in various combinations,[11,12] to additional biochemical or ultrasound-based risk factors.[13,14] Existing screening strategies for FGR are not effective. Many do not differentiate between early and late-onset FGR, or with SGA fetuses.[15,16] A Cochrane review of randomised trials on universal screening with ultrasound in pregnant women compared with a strategy of selective screening in high-risk women for FGR did not show any reductions in perinatal mortality and morbidity.[17] The latter strategy detects only 20% of small babies, while with the former strategy, two otherwise normal small babies are picked up for every SGA fetus with complications identified.[18] Universal ultrasound screening of all women for detection of FGR can significantly strain finite resources. Implementation of such a strategy in low-risk women in France did not lower the rates of complications in SGA fetuses, but resulted in iatrogenic prematurity in screen-positive pregnancies.[19] Similarly, a cluster randomised trial comparing routine ultrasonography in third trimester to usual care of clinically indicated ultrasonography, showed only a moderate increase in the detection of SGA infants, but with increases in induction of labour, and no reduction in severe adverse perinatal outcomes in low-risk pregnancies.[20] The National Institute for Health and Care Excellence's (NICE) antenatal care guideline concluded that 'the methods by which an SGA fetus can be identified antenatally are poorly developed or are not tested by rigorous methodology'.[21]

Numerous primary studies and aggregate meta-analyses have reported on the accuracy of individual clinical, biochemical and ultrasound markers or multivariable prediction models to predict either FGR or SGA fetus. Although more than 20 prediction models were developed, none were recommended for use in routine clinical practice.[22–26] This is due to difficulties involving the design, population, tests and outcomes of existing research to predict, screen or detect FGR. Firstly, the terms 'prediction' and 'screening', which have separate objectives, are often used interchangeably.[27] In the former, the outcome of interest (FGR) has not yet occurred, while in the latter, the focus is on accurately detecting established

FGR. Some of the models to predict FGR use tests as late as 36 weeks of pregnancy, which are more relevant for diagnosis than prediction.[26] Secondly, the population studied is often only limited to specific subgroups such as nulliparous women.[18] Thirdly, the predictors have often been dichotomised, thereby reducing their power. Fourthly, before they can be recommended for use in clinical practice, the predictive performance of prediction models needs to be appropriately evaluated in populations in which it is intended for use, and external to that used to develop the model. Fifthly, studies often predict SGA rather than FGR infants. FGR is variously defined using either ultrasound characteristics [EFW, fetal abdominal circumference (AC), Doppler blood flows] or by using birthweight.[28] Furthermore, both EFW and birthweight have been reported in centiles that were either adjusted for various maternal characteristics (customised) or for only gestational age (population based),[29,30] additionally, the centile cut-offs to define growth restriction are varied (<10th, <5th, <3rd).

Meta-analysis of individual participant data (IPD), where the raw participant-level information is obtained and synthesised across multiple datasets can help overcome the above limitations.[31–35] Availability of the raw data from multiple datasets will substantially increase the sample size beyond what is achievable in a single study. It will allow the standardisation of the definition of FGR and predictors across datasets and enables assessment of differential accuracy of prediction models in different subgroups of women across a range of clinical settings. IPD meta-analysis enables the evaluation of multivariable models that contain multiple candidate predictor variables, it allows for methods that directly handle missing predictor and outcome data, allows the examination and accounting of heterogeneity (e.g. in baseline risks), and can develop, validate and tailor the use of the most accurate prediction models to the appropriate population.

We have previously established the International Prediction of Pregnancy Complications (IPPIC) network of global researchers,[36] with access to IPD from over three million pregnancies and undertook an IPD meta-analysis to accurately identify fetuses at risk of growth restriction and perinatal complications, to predict the extent of smallness using prediction models, and also to assess the relative costs and outcomes of a strategy of predicting FGR using any newly developed prediction model.

Chapter 2 Objectives

We aimed to identify and externally validate existing prediction models for FGR and birthweight, and then if necessary, update or develop and validate further prediction models in pregnant women to determine (1) the overall risk of delivering a growth-restricted fetus (birthweight <10th centile adjusted for gestational age, with serious perinatal complications of stillbirth, neonatal death or delivery before 32 weeks); and (2) the birthweight if delivered at various gestational ages (with flexibility to convert into centiles using existing fetal growth standards) to assess the extent of smallness, using data from the large IPPIC IPD repository.

Primary

1. To establish whether existing prediction models for FGR and birthweight are suitable for the target population or if new models are needed through external validation, and where possible, recalibration of existing prediction models.
2. Using IPD meta-analysis, to develop and validate [using internal-external cross-validation (IECV)] new multivariable prediction models for (i) FGR (SGA with serious perinatal complications) (IPPIC-FGR Model 1); and (ii) birthweight at various potential gestational ages at delivery (IPPIC-birthweight Model 2) based on:

   – clinical characteristics only
   – clinical and biochemical markers
   – clinical and ultrasound markers
   – clinical, biochemical and ultrasound markers.

Secondary

1. To compare the predictive performance of models according to (i) population (selected – high/low risk; unselected); (ii) trimester of testing (first <14 weeks; second ~20 weeks; third ~28 weeks); (iii) choice of predictors (clinical only; clinical and ultrasound; clinical and biochemical; clinical, ultrasound and biochemical; and (iv) onset of FGR (early <32 weeks; late ≥32 weeks).
2 To assess if the performance of the prediction models is generalisable for various definitions of FGR such as (i) ultrasound parameters determined by Delphi consensus;[37] and (ii) birthweight <10th centile adjusted for gestational age with associated neonatal morbidity,[38] and assess the association between various birthweight centiles (<10th, <5th, <3rd centiles) calculated using (i) customised and (ii) population-based standards, and perinatal mortality and morbidity.
3. To examine the clinical utility of the prediction models using decision curve analysis (DCA).
4. To assess the costs and outcomes and the potential impact of resource use of the prediction models.

Chapter 3 Methods

Our IPD meta-analysis followed existing recommendations for prediction model development and validation,[39–42] and used a prospective protocol registered with International prospective register of systematic reviews (CRD42019135045). Our reporting utilises the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis and Preferred Reporting Items for Systematic Reviews and Meta-Analyses-IPD reporting guidelines for prediction models and IPD meta-analysis.[27,43]

The International Prediction of Pregnancy Complications Network

The IPPIC Network database is a living data repository of IPD from pregnant women. Methods on how cohorts within the IPPIC Network database were identified and harmonised have been described in detail in our earlier publications.[36,44,45] Briefly, cohorts within the database were identified through a systematic search to identify primary studies reporting risk factors for pregnancy complications including pre-eclampsia (PE), stillbirth and FGR.[46] Authors of relevant studies were invited to join the network and share their primary IPD in any format, along with data dictionaries or descriptions. The data were deposited in a custom-built database, formatted, cleaned and harmonised, and the quality of each cohort and its IPD was assessed using the following domains of the Prediction study Risk of Bias Assessment (PROBAST) Tool: participants (adequate description of data sources, details on recruitment), predictors (appropriately defined, assessed blinded to outcome, assessed in the same way for all participants) and outcome (appropriately defined and determined in a similar way for all participants, predictors excluded from the outcome definition, outcome determined without knowledge of predictor information and appropriate interval between assessment of predictor and outcome determination).[47]

The IPPIC-IPD included data from large birth or population-based cohorts, registry data, unpublished data from hospital records or study cohort data. Study population varied from low to high risk of development of complications. The predictor variables harmonised within the IPPIC Network repository are those that are easy to obtain in a clinical setting, as agreed by the collaborative group.[44] The network currently includes more than 150 collaborators from 26 countries, contributing IPD from over 3 million pregnancies, reporting maternal characteristics, obstetric history, clinical assessment and tests, as well as various maternal and offspring outcomes. Cohorts that addressed the structured question in *Table 1* were considered for inclusion in the IPD meta-analysis.

Primary outcomes

The primary outcomes were (1) FGR (birthweight <10th centile adjusted for gestational age, with stillbirth, neonatal death or delivery before 32 weeks); and (2) birthweight for deliveries at various gestational ages to reflect the extent of the restricted growth.

Rationale for the choice of outcomes

Fetal growth restriction: FGR defined as birthweight <10th centile adjusted for gestational age, with severe complications was chosen for the following reasons: the definition excludes small but healthy babies; the components of the composite include severe complications of mortality or extreme prematurity (both iatrogenic and spontaneous preterm births before 32 weeks are reflective of the severity of the condition). Any prediction model will need to take into consideration the effects of treatment paradox, where delivery could have prevented stillbirth or neonatal death that may have otherwise occurred.[48] This was addressed by including delivery before 32 weeks as a component of the outcome. Birthweight centiles were calculated based on published ranges of birthweights for live births

TABLE 1 Structured questions for IPD meta-analysis on prediction of birthweight and FGR with complications

Population	Pregnant women
Predictors	*Maternal clinical characteristics: Maternal characteristics:* Age, BMI, smoking, alcohol or substance misuse, exercise, diet; *Medical history:* chronic hypertension, diabetes, renal disease, heritable thrombophilia, autoimmune disease, cardiac disease; *Obstetric history:* parity, previous SGA, previous stillbirth, previous PE, pregnancy interval; *Current pregnancy:* mode of conception, weight gain, early pregnancy bleeding *Biochemical markers:* PlGF, PAPP-A, sFlt-1, AFP, HCG, urine dipstick, 24-hour urine protein *Ultrasound markers:* Uterine artery Doppler (RI, PI, unilateral or bilateral notching), AC, fetal, CPR, EFW, fetal echogenic bowel, NT
Outcomes	**Primary outcomes:** FGR defined as birthweight <10th centile adjusted for gestational age at delivery, complicated by stillbirth or neonatal death or delivery before 32 weeks; birthweight at various gestational ages **Secondary outcomes:** Early onset (<32 weeks) and late-onset (≥32 weeks) FGR Ultrasound-based diagnosis for early (EFW <3rd centile, AC <3rd centile, absent end diastolic flow in umbilical artery Doppler) and late FGR (EFW <3rd centile, AC <3rd centile) Neonatal morbidity: cord blood pH < 7 at birth, hypoxic-ischemic encephalopathy, respiratory distress syndrome, septicaemia, admission to neonatal intensive care unit, Apgar score < 7 at 1' and 5'
Study design	IPD meta-analysis of observational studies and cohorts nested within randomised trials

AFP, alpha-fetoprotein; BMI, body mass index; CPR, cerebral-placental ratio; HCG, human chorionic gonadotropin; NT, nuchal translucency; PAPP-A, pregnancy-associated plasma protein A; PI, pulsatility index; PlGF, placental growth factor; RI, resistance index; sFlT-1, soluble fms-like tyrosine kinase-1.

from King's College Hospital, London, between March 2006 and October 2015.[49] We applied the normal ranges of birthweights according to gestational age to determine the birthweight centile in the IPD.

Birthweight: Existing prediction models use arbitrary cut-offs to define FGR or SGA fetus using only birthweight <10th or <3rd centile. Dichotomisation of the outcome limits the power and usefulness of a prediction model. Besides, the prognosis for a fetus with a predicted birthweight on the 3rd centile at 26 weeks is far worse than that predicted to be on the 9th centile at 37 weeks, despite both being labelled as small with <10th centile birthweight. A baby diagnosed to be small using a particular fetal growth standard (e.g. GROW, INTERGROWTH 21st, WHO)[50,51] may not be categorised as so with another standard, thereby limiting the generalisability of the model. To address this, we used birthweight as our outcome to be predicted at various potential gestational ages at delivery for the following reasons: (1) it is a continuous measure not limited by arbitrary cut-offs; (2) the predicted birthweight can be converted into predicted centiles using any fetal growth standard in use; and (3) it provides information on both severity of the restricted growth, and the expected timing of onset to plan appropriate management. For example, a baby with a predicted birthweight on the 5th centile at 28 weeks' gestation will require frequent monitoring starting from 26 weeks.

Updating literature search

Existing prediction models for fetal growth restriction

We updated our previous literature search (search to July 2012)[26] to identify additional models for FGR or birthweight published up to August 2019. We searched MEDLINE and EMBASE databases without any language. We included studies reporting multivariable (at least three variables) models on the risk of FGR (birthweight <10th centile adjusted for gestational age, with severe complications of either stillbirth, delivery before 32 weeks or neonatal death at any time) for use in early pregnancy (≤28 weeks' gestation) or birthweight. We excluded studies of models that predicted FGR as part of

any other combinations of composite adverse outcomes, contained predictors that were not measured in any of the cohorts within the IPPIC IPD, or did not publish the reported model equation (including model intercept). Two independent researchers undertook study selection and data extraction, with disagreements resolved by discussion.

Strengthening the IPPIC Network

We augmented the existing live IPPIC data repository by including additional datasets from studies providing relevant data to predict FGR or birthweight, based on our previously conducted systematic reviews.[46,52] The systematic review methods have been published elsewhere. Briefly, we searched MEDLINE, EMBASE, Cochrane (Wiley) CENTRAL, Science Citation Index (Web of Science), CINAHL (EBSCO), ISRCTN Registry, UK Clinical Trials Gateway, WHO International Clinical Trials Portal and ClinicalTrials.gov; specialist abstract and conference proceeding resources (British Library's ZETOC and Web of Science Conference Proceedings Citation Index) to identify relevant studies. We adhered to PRISMA guidelines on reporting, and the reviews were based on prospective protocols. Two reviewers independently screened abstracts, extracted data and carried out quality assessment. We invited authors of all primary studies identified from the reviews to join the IPPIC Network and share their IPD, with at least two further email reminders if no response was received. We additionally invited investigators of primary studies or datasets not included in the reviews but identified through our links with other collaborative groups, if they contained relevant information needed (see *Table 1*).

We standardised the data that were shared by recoding and harmonising them in line with the clean formatted IPPIC datasets. We undertook rigorous range and consistency checks using methods detailed in *The International Prediction of Pregnancy Complications Network* above and previous publications.[36,44,45] We continued to contact authors to share their data until the July 2020 deadline for receiving new datasets. We set the deadline to allow time for cleaning and formatting of the data prior to analysis. Any IPD shared beyond this time period was not included in our analysis. All relevant data available in the IPPIC repository at the time of database-lock on 31 January 2020 were included for external validation of existing prediction models; we included data in the repository by 31 August 2020 to develop the IPPIC prediction models.

Prioritisation of predictors

We carried out a prospective two-round e-survey of IPPIC Network collaborators, to prioritise the most clinically relevant predictors of FGR, to be considered in the development of the prediction models. Predictors were identified from existing systematic reviews.[22] The first round of the survey included explanation of the study and consent process, followed by a list of predictors identified from the systematic reviews. Collaborators were asked to rank the importance of each predictor variables identified on a scale from 1 (not important) to 5 (very important). The predictors were classified as 'consensus in' if ≥70% of responders gave a score of 4 or 5 and <15% score 1 or 2, or 'consensus out' if ≤50% of responders gave a score of 4 or 5 and ≤30% gave a score of 3. Anything else was classified as 'no consensus'.

In the second round of the prioritisation survey, collaborators were invited to a video Zoom conference on 28 July 2020 and asked to reassess predictors ranked as 'consensus in' or 'no consensus' from the first round of voting. An open discussion took place on each outcome and collaborators were encouraged to consider how important the measurement of each predictor was as a predictor of FGR. The Zoom polling function was used to vote on a scale from 1 (not important) to 5 (very important) for each predictor and analysed using the same method in the round one survey. Any variable still classed as 'no consensus' was discussed at the meeting and a final classification agreed upon.

Sample size considerations

The effective sample size for the development and validation of prediction models is driven mainly by the total number of events (for logistic regression of a binary outcome) or the total number of subjects (for linear regression of a continuous outcome). To reduce the potential for overfitting and optimism during model development, the number of subjects/events must be large relative to the number of candidate predictor parameters to be considered for inclusion in the model.

For the external validation of published prediction models, sample size calculations aim for precise estimates of the predictive performance,[53–55] and suggest at least 100 events and 100 non-events for binary outcomes, which we hoped to meet – though again our sample size was fixed, based on the IPD available that recorded the required predictors available for each model.[54]

The IPPIC-FGR model to be developed has the binary outcome of FGR (birthweight <10th centile adjusted for gestational age, with serious perinatal complications). For this, Riley *et al*. proposed sample size calculations to ensure small optimism in predictor effect estimates, a small difference in the apparent and adjusted estimates of Nagelkerke's R^2, and precise estimation of the overall risk in the population.[56] For example, based on an estimated FGR prevalence of 0.73%, with a maximum possible Cox–Snell R^2_{CS} of 0.08, and an assumed lower bound for the apparent Nagelkerke's R^2_N of 0.32 based on previously published models,[57] a minimum sample size of 34,906 women with 255 FGR events is required to meet the criteria when considering up to 50 predictor parameters. As our sample size was fixed (as it is dependent on the available IPD), for the models developed we restricted the number of candidate predictor parameters below 50 so that our sample size would easily meet the criteria of Riley *et al*.

The IPPIC-birthweight model has the continuous outcome of birthweight. Riley *et al*. further recommend that the sample size used to develop such a model should be sufficient to ensure small optimism in predictor effect estimates, a small difference between the apparent and adjusted R^2, precise estimation of the mean predicted birthweight (the model intercept), and precise estimation of the model's residual standard deviation (SD).[58] For example, assuming a lower bound for the anticipated adjusted R^2 of 0.5 in the new model, and an intercept value of −0.935 with standard error 0.043 (on the $\log_{10}$ scale) based on previous literature,[59] a minimum sample of 618 women is required to consider up to 50 predictor parameters. Again, our sample size was fixed according to the IPD available, and we restricted the number of candidate predictor parameters to below 50, in order to meet the criteria by Riley *et al*.

Data synthesis

We used SPSS Version 27 (IBM SPSS Statistics for Windows) to analyse the Delphi survey findings that prioritised the predictors of FGR. All other analyses were carried out using Stata MP Version 16.

External validation of existing prediction models

Prediction models were validated if at least one IPPIC-IPD cohort contained all the predictors included in the model, and the model outcome occurred in some of the participants in the IPD cohort. We did not exclude women with multifetal pregnancies (i.e. twins/triplets) from our analysis. Women may have become pregnant multiple times during the course of data collection in an IPD cohort, and each pregnancy was considered as a distinct observation for validation. Although two or more pregnancy outcomes from the same women are likely to be correlated, the number of women with consecutive pregnancies is small relative to the total number of pregnancies contained in the IPD database. Furthermore, our external validation aims to confirm whether these prediction models are accurate for all potential applications, regardless of whether they have been applied to the same women previously.

Missing data

Partially missing predictor and outcome variables were imputed using multiple imputation by chained equations, assuming that individual values were missing at random (MAR). Imputation was performed separately for each cohort to allow for the clustering of individuals within each cohort. The number of imputed datasets (*m*) was set equal to the largest percentage of incomplete observations in any of the individual studies, with the same *m* being used for imputation in all studies.[60] Rubin's rules were then applied to combine estimates across imputations.[61]

All predictors and outcomes were included in the imputation models, to help ensure the MAR assumption was more reliable. Linear regression was used to impute for approximately normally distributed continuous variables, and logistic regression was used for binary variables. Predictive Mean Matching was used for the imputation of categorical variables, due to convergence issues with multinomial logistic regression within Stata's mi impute command. Where a study had participants with missing outcome values, these outcomes were imputed in the same way as missing predictor values (depending on data type). Observations with imputed outcomes were then deleted prior to analysis. Imputed values were then checked through visual inspection of histograms (continuous variables) and tables (categorical variables) to ensure values were realistic and consistent across imputed data sets. Complete case analyses were also performed for comparison.

Assessment of model performance

Within each cohort, the model equations were applied to each participant in the IPD to calculate the linear predictor (LP) value for that individual ($LP_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots$, the value of the linear combination of predictors in the model equation for individual *i*). For models predicting the continuous outcome of birthweight, the final prediction was equal to LP_i for each individual. For models predicting the binary outcome of FGR (meeting the requirements of our definition), the probability of FGR for a pregnancy was calculated as $p_i = \frac{e^{LP_i}}{1+e^{LP_i}}$. We then summarised the distribution of predictions by cohort using histograms and by determining the median and interquartile range (IQR).

The predictive performance of each model was evaluated using measures of calibration, referring to how well the predictions from the model agree with the observed outcomes,[62,63] and discrimination, referring to how well the model differentiates between those who have the event and those who do not (only for binary outcome models).

Calibration was assessed across the entire population, as well as in subgroups according to gestational age at delivery (<28 weeks, 28–31 weeks, 32–36 weeks and ≥37 weeks) to assess differential model performance in these populations. Calibration was assessed using two measures:

1. The calibration slope, which is the slope of the regression line fitted between the observed and predicted outcomes on the original scale for continuous outcomes ($Y_{TRUEi} = \alpha + \beta(LP_i)$, where β is the estimated calibration slope) and on the logit scale for binary outcomes (logit-p = $\alpha + \beta(LP_i)$). Ideally, the calibration slope should be equal or very close to 1.
2. Calibration-in-the-large (CITL), which indicates the extent to which model predictions are systematically too low or too high across the dataset and should ideally be equal to 0. The estimate of CITL was obtained from α when fitting the above calibration model with $\beta = 1$.

We also produced calibration plots plotting the observed (O) against the expected (E) birthweight value for each patient (continuous models), or observed versus expected FGR probabilities across risk groups (binary models). As calibration plots cannot be pooled across imputations, plots were assessed separately for each imputed dataset.[64] Where performance looked similar across imputations, a calibration plot was presented using predicted outcome values that were pooled across imputed datasets for each individual outcome. A LOWESS smoother was applied to each plot to show the non-linear calibration slope, calculated using all participants (avoiding risk grouping), across the entire

range of risk predictions. Calibration plots are presented with a diagonal line to show perfect calibration (where observed exactly equals expected), and close proximity of points to this line can be interpreted as good calibration performance of the model. Points lying above the diagonal indicate predictions that are lower than observed outcomes (underprediction), while points lying below the diagonal show where predictions are higher than observed outcomes (overprediction).

Discriminative ability of a binary outcome model was assessed using the *c*-statistic, (equivalent to the area under the receiver operating characteristics curve, with a value of 1 indicating perfect discrimination and 0.5 indicating no discrimination beyond chance). For each model validation, predictive performance measures were summarised across the cohorts using a two-stage IPD meta-analysis approach.

Validation performance measures were first calculated for each cohort separately and then pooled[65] using a random-effects meta-analysis model estimated using restricted maximum likelihood estimation (DerSimonian–Laird estimates used for subgroup analysis). Random-effects meta-analysis was used as we assumed that the performance of a model would differ across populations, due to case-mix variation.[31,66] Random-effects meta-analysis also allowed us to quantify heterogeneity in predictive performance across cohorts and to predict model performance in other similar settings using approximate 95% prediction intervals.[67] The calibration slope and CITL were pooled on their original scale, while the *c*-statistic was pooled on the logit scale[68] with the standard errors of logit-C calculated using the delta method.[66] Model performance was summarised for each predictive performance statistic as the average and 95% confidence interval (CI) for the average performance statistic. CIs were derived using the Hartung–Knapp–Sidik–Jonkman variance correction, to account for uncertainty in the between-study variance (often due to few studies being present in the meta-analysis).[69]

We summarised the heterogeneity in model performance across cohorts τ^2, with approximate 95% prediction intervals calculated using the approach of Higgins *et al.*[70] We showed the model performance across cohorts graphically using forest plots for each predictive performance measure, and scatter plots to show both calibration measures in combination (CITL and calibration slope, to give an impression of the overall calibration performance of the model).

Decision curve analysis

We assessed the clinical utility of a model for predicting FGR (binary outcome) using DCA.[71,72] The net benefit (NB) of the model, weighing up potential benefits and harms was plotted at various clinically relevant threshold probabilities. For a probability threshold (*pt*), the NB was calculated as $\frac{TP}{N} - (\frac{FP}{N} \times \frac{p_t}{1-p_t})$, where *TP* and *FP* represent the numbers of individuals with a predicted probability ≥ p_t that do and do not have FGR, respectively, and *N* is the total sample size.[71,72] The model with the greatest NB for a particular threshold is considered to have the most clinical value. Threshold probabilities refer to cut-off points, where in practice a prediction greater than the threshold would be treated as 'high risk', and a prediction below this threshold would be considered low risk. Decision curves were compared between models and against 'treat-all' and 'treat-none' strategies (where an intervention is for everyone and no one, respectively), focusing across the range of threshold probabilities at which the model may be clinically useful. Based on clinical discussion, the threshold range was agreed in advance to be 0.01 to 0.2, meaning predicted FGR risks in the range from 1% to 20% were considered as potential cut-points for informing changes to treatment in practice, and so a NB in this range was desired.

To assess the clinical implications of using linear regression models for predicting birthweight to imply FGR risk, predicted probabilities were gained from the outcome of the linear regression model using the distribution of the predicted values across individuals *i* (Y_{PREDi}), where Y_{PREDi} was assumed to follow a student's t distribution with $n - p - 1$ degrees of freedom (*p* denoting the number of predictor parameters in the prediction model and *n* the number of participants).

Recalibration of existing fetal growth restriction prediction models

Where the existing models were miscalibrated, recalibration methods were considered. In particular, the intercept and slope of the LP were to be re-estimated to improve CITL and the calibration slope.

Development and validation of new or updated prediction models

To develop and validate new prediction models for (1) FGR (birthweight <10th centile adjusted for gestational age, with stillbirth or neonatal death or delivery before 32 weeks) and (2) birthweight at various gestational ages, we considered cohorts contained within the data repository at final database lock in August 2020.

Candidate predictors for model development were informed by predictors included in existing prediction models and by clinical experts in the collaborative group as detailed in *Prioritisation of predictors*. Our aim was to produce predictions conditional on assumed gestational ages at delivery, and therefore gestational age at delivery was included as a predictor in our models. Although the actual gestational age at delivery would not be available at the moment of prediction, producing the models in this way allows a range of assumed gestational ages at delivery to be entered for each woman, and a graph of predictions against gestational age to be made for them, to give a more complete picture over time. Example plots of such predictions are given later in the report.

To select datasets to use for development of a new prediction model, it was necessary to compromise between the number of datasets included and the potential predictors that could be considered for inclusion in the models (as not all predictors were available in all datasets). The aim was to do this in such a way to maximise both. We undertook the following process:

1. Summarised the number of datasets, total sample size and number of events available for each candidate predictor considered for inclusion.
2. Ranked the prioritised predictors based on the number of cohorts reporting the predictor in the IPD.
3. Started with the most commonly reported predictor and added prioritised predictors in a sequential manner to obtain the set of predictors which maximised the number of cohorts in the IPD, number of participants and number of events.
4. Stopped when adding any further predictors resulted in a sizeable loss of cohorts, participants or events and ensuring there was sufficient data to meet the sample size criteria set out in *Sample size considerations*.

Missing data

The number and proportion of missing values for each potential predictor and outcome were summarised by cohorts. Predictors were considered to be systematically missing for a cohort if they were not recorded for any or were recorded for very few individuals (<10%) in that cohort. Predictor values were not imputed for any cohort in which they were systematically missing.

Multiple imputation was implemented in each cohort separately to acknowledge the clustering of individuals within, and to retain heterogeneity between, cohorts.[35] We generated 100 imputed datasets (to exceed the largest percentage of incomplete observations in any of the individual cohorts), using chained equations, for each IPD cohort with any partially missing candidate predictors or outcome variables. Continuous variables were imputed using linear regression, binary variables were imputed using logistic regression and categorical variables were imputed using predictive mean matching. Complete predictors were also included in the imputation models as auxiliary variables. The imputation model included all candidate predictors and both outcome variables (birthweight and FGR).

Due to the difficulties in handling non-linearity in model development, and accounting for different non-linear functions in the imputation, a pragmatic decision was made to perform a preliminary complete

case analysis to look for potential non-linear relationships between continuous candidate predictors and each outcome variable using multivariable fractional polynomial (MFP) models. Visual comparison of FP1 and FP2 functions was used to decide on the complexity of the functions to be included. If there was little difference between the shape of FP1 and FP2 functions, the simpler FP1 function was selected. Where a non-linear function was selected for a variable in the complete case analysis, rather than assuming that the FP1 function selected was correct, we included each of the possible (FP) functions in the imputation model, to enable this non-linearity to be considered during model development.

After imputation, the distributions of values for imputed variables were checked by plotting the mean ± SD for continuous variables against the imputation number (including the original unimputed data, imputation 0, for reference). For categorical variables, the proportions in each category were compared across imputations and to the original unimputed data.

Model development and variable selection

Prediction models were developed using random intercept regression models with backward elimination for variable selection. The random intercept was used to account for clustering of women within the individual cohorts.

Variable selection and consideration of the functional form for continuous variables took place within each cycle of the IECV (detailed below). An MFP approach was used, in which fractional polynomial functions were tested for each continuous variable (identified in the previous complete case analysis to potentially have a non-linear association with the outcome) to determine the 'best' functional form of that variable in the multivariable model (i.e. in the presence of all other variables).

At each stage of the variable selection process, the same model (i.e. including the same candidate predictors) was fitted to all imputations, and pooled Wald tests (using Rubin's rules) were used for backward elimination, with $p > 0.157$ (proxy for AIC) for exclusion.[73]

Heuristic shrinkage was calculated following the method proposed by Van Houwelingen and le Cessie[74] for the final model in each imputation and pooled across imputations using Rubin's rules to obtain the average shrinkage factor. This average shrinkage factor was then applied to each beta coefficient in the models, and subsequently average intercept values were re-estimated (holding fixed the shrunken beta coefficients) to ensure predictions in each dataset were correct on average.

Internal-external cross-validation

An IECV approach was used for validation, as IPD were available from multiple cohorts.[31,32] Using this approach, a model is developed using all but one cohort which is reserved for 'external' validation. The model is then internally validated using the same data, using methods such as bootstrapping to calculate the optimism in the model performance and the shrinkage factor. Bootstrapping was not practical computationally given the need to incorporate both non-linear trend examinations, variable selection and multiple imputation. Therefore, an approximate heuristic shrinkage factor was calculated (not accounting for the variable selection process) following the method proposed by van Houwelingen and le Cessie[74] and applied to the regression coefficients as described above.[65]

Following shrinkage, the model's average intercepts were re-estimated to ensure predictions were correct on average. This then provided the 'shrunken' model equation. This 'shrunken' model was then applied to the omitted study to calculate the predicted birthweight at the observed gestational age at delivery, and then the predictive performance measures were calculated using CITL, the calibration slope, the *c*-statistic and Nagelkerke's R^2 (as described in *Assessment of model performance*). This completes one cycle of IECV, and the process was repeated multiple times, each time reserving a different study for 'external' validation (see *Figure 1*). Calibration plots were also produced for each cycle

of IECV, plotting average observed and expected values across imputations (where imputation-specific calibration plots were consistent with one another).

Following IECV, there were multiple values for each predictive performance measure (one from each cohort). These estimates were summarised using random-effects meta-analysis to give a pooled estimate of overall performance on IECV. Apparent performance of the model was also calculated for each cohort individually (and across the full dataset) using the average intercept term, to better approximate how the model would be applied in new individuals. Cohort-specific apparent predictive performance was also summarised across cohorts using random-effects meta-analysis to give a pooled estimate of overall apparent model performance.

For these random-effects meta-analyses, the calibration slope and CITL were pooled on their original scale, while the *c*-statistic was pooled on the logit scale[68] with the standard errors of logit-C calculated using the delta method.[66] CIs were derived using the Hartung–Knapp–Sidik–Jonkman variance correction.[69]

Decision curve analysis

For the binary outcome model, decision curves were produced (as described in *Decision curve analysis*) within each study cohort individually, as well as within the full dataset used for development. Expected numbers of true/false positives (T/FP) and true/false negatives (T/FN) per 1000 women based on using the model are also reported for a selection of potentially clinically relevant threshold probabilities, along with estimates of sensitivity and specificity for the model at each threshold, with the region between thresholds of 1% to 20% of most interest.

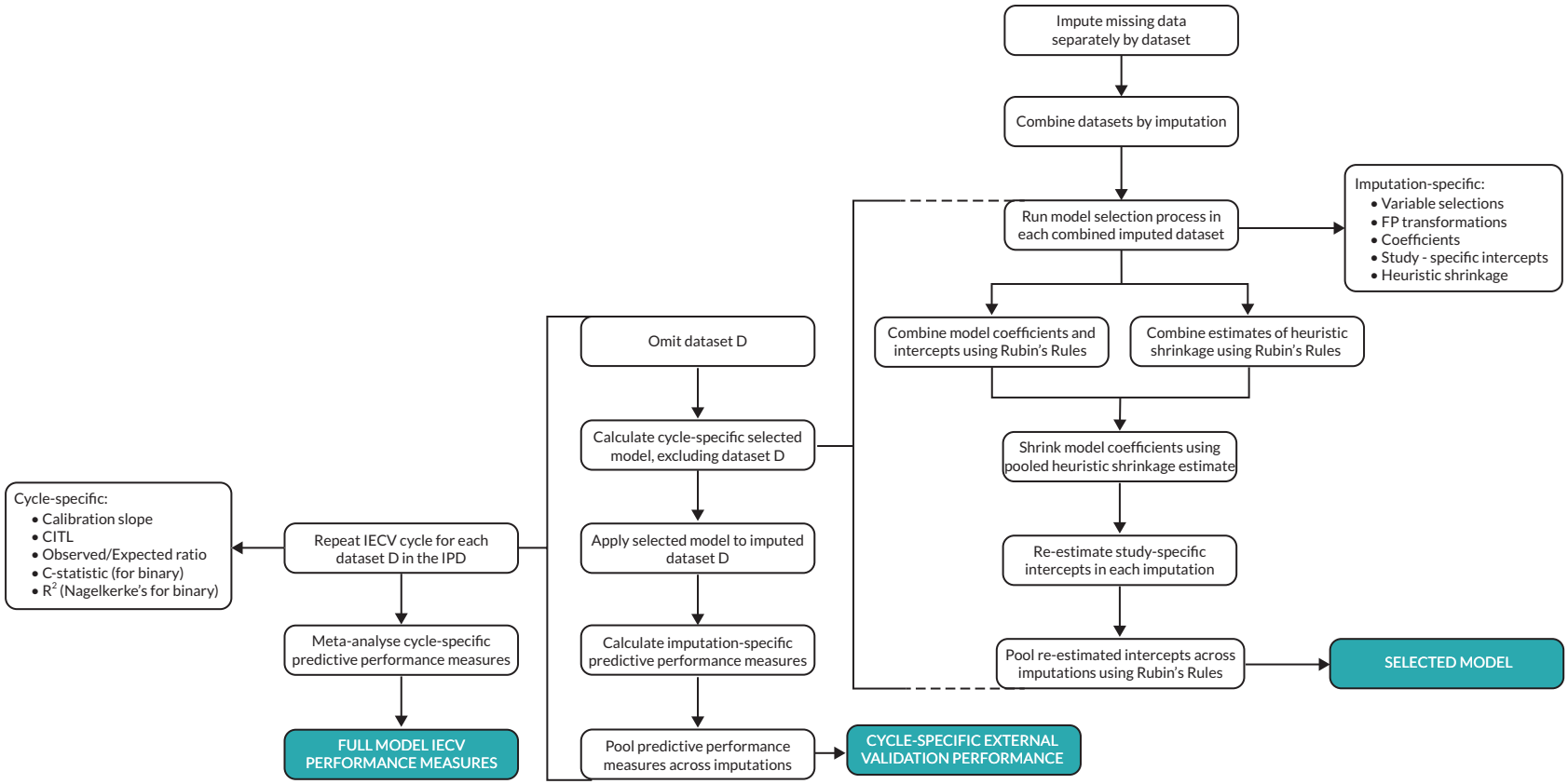

FIGURE 1 Flow diagram showing processes involved in development and validation of the prediction model.

Chapter 4 Characteristics of IPPIC cohorts and prioritisation of candidate predictors for model development

Characteristics of IPPIC cohorts

Overall, 94 cohorts were available in the IPPIC data repository (including 16 added cohorts), contributing data from 4,539,640 pregnancies.[18,75–164] About half the studies in the repository were prospective cohort studies (57%, 54/94), 16% (15/94) were randomised trials and 14% (13/94) were prospective registry datasets or birth cohorts. One dataset was an IPD of 31 RCTs. Most of the datasets included pregnant women from Europe (61%, 57/94), 16% (15/94) from North America, 6% (6/94) from South America, 6% (6/94) from Asia and Australia and 1 from Africa. Five datasets provided included participants from multiple countries such as Argentina, Colombia, Kenya, India, Peru, Thailand, Vietnam, Lebanon, Mexico, Mongolia, Uganda, Nigeria and New Zealand. About a quarter of datasets received were on women with high-risk pregnancies only (26%, 24/94), 13% (12/94) on low-risk pregnancies and more than half (61%, 57/94) included women with any risk pregnancies. Detailed study characteristics of all IPPIC datasets are provided in *Appendix 1*.

Prioritisation of candidate predictors of fetal growth restriction: Delphi survey findings

Forty collaborators participated in the first round of the e-survey. Most of the participants were from Europe (65%, 26/40), five each from the American and Oceania continents, three from Asia and one from Africa. Twenty-three participants took part in the second round of the prioritisation survey which took place via a Zoom video conference. Thirteen participants were from Europe (57%, 13/23), five from America (22%, 5/23), three from Oceania (3/23) and one each from Asia and African continents.

We identified 33 predictor variables from existing systematic reviews (18 clinical characteristics, 7 biochemical markers and 8 ultrasound markers). Additionally, between the first and second round of the survey, our external validation of existing prediction models for FGR identified a promising model with reasonable performance (see *Chapter 5*).[59] It was decided to take forward all predictors in the model as candidate predictors in our model development. These predictors were therefore included as candidate factors regardless of the ranking obtained from the first round of voting, and they were not considered by collaborators during the second round of voting.

The predictors included from the Poon 2011 prediction model, as well as predictors voted in/out following the two-round survey are provided in *Table 2*. A comparison of possible sample sizes based on combinations of candidate predictors in addition to the predictor variables from the Poon 2011 model[59] was conducted and yielded the below (see *Table 3*). At each stage, the additional variable that maximised the number of cohorts, participants and FGR events was carried forward with the variables already selected. The process was then repeated considering the other candidate variables in the next iteration. In *Table 3*, bold text shows which variable was carried through to the next iteration, while red italics shows a variable was removed at that point, as only one study measured that combination of variables.

The final list of candidate predictors included those from the Poon 2011 model,[59] along with previous PE, previous stillbirth and having had a previous SGA baby. This combination of predictors resulted in a restriction of analysis to 4 cohorts with 237,228 pregnancies and 1729 events (which met the sample size requirements discussed in *Sample size considerations*) for model development.

TABLE 2 Predictors of FGR prioritised in Delphi survey

Potential candidate predictors	Excluded as not prioritised by researchers
Included from existing Poon 2011 model[59]	
Gestational age	Vaginal bleeding in this pregnancy
Mother's weight	Pregnancy interval
Mother's height	Alcohol intake
Mother's age	Drug misuse
Parous	Chronic kidney disease
Smoking	History of autoimmune disease
Ethnicity (white, black, Asian, Hispanic, mixed or other)	History of heritable thrombophilia
Chronic hypertension	History of cardiovascular disease
Diabetes	BMI
Assisted conception	Uterine artery Doppler notching
	Uterine artery Doppler raised RI
From prioritisation by collaborators	Suboptimal fetal growth by AC centile
Previous stillbirth	Fetal CPR
Previous SGA baby	Fetal echogenic bowel
Previous history of PE	NT
PlGF	HCG
Uterine artery Doppler raised PI	AFP
EFW	PAPP-A
	sFlt-1
	Proteinuria – urine dipstick > 2 + protein
	Proteinuria – >300 mg/24 hour collection

AFP, alpha-fetoprotein; BMI, body mass index; CPR, cerebral-placental ratio; HCG, human chorionic gonadotropin; PAPP-A, pregnancy-associated plasma protein A; PI, pulsatility index; PlGF, placental growth factor; RI, resistance index; sFlT-1, soluble fms-like tyrosine kinase-1.

TABLE 3 Candidate predictors for IPPIC-FGR model finalised based on data availability, existing literature and clinical consensus

Root	Addition	Number of datasets	Number of participants	Number of events
Poon 2011 predictors	Previous stillbirth	10	674,529	6394
	Previous SGA baby	5	238,428	1743
	Previous PE	**11**	**677,370**	**6433**
	PlGF	6	12,436	61
	Uterine artery Doppler PI	6	12,436	61
	Uterine artery Doppler PI (T1)	5	8224	49
	Uterine artery Doppler PI (T2)	5	17,917	45

TABLE 3 Candidate predictors for IPPIC-FGR model finalised based on data availability, existing literature and clinical consensus (*continued*)

Root	Addition	Number of datasets	Number of participants	Number of events
	Uterine artery Doppler PI (T3)	*1*	*8824*	*12*
	EFW	4	247,342	1733
Poon 2011 predictors + previous PE	**Previous stillbirth**	**8**	**670,254**	**6384**
	Previous SGA baby	5	238,428	1743
	PlGF	5	16,985	63
	Uterine artery Doppler PI	4	8161	51
	Uterine artery Doppler PI (T1)	3	3949	39
	Uterine artery Doppler PI (T2)	4	15,281	40
	EFW	4	247,342	1733
Poon 2011 predictors + previous PE + previous stillbirth	**Previous SGA baby**	**4**	**237,228**	**1729**
	PlGF	2	9869	14
	Uterine artery Doppler PI	*1*	*1045*	*2*
	Uterine artery Doppler PI (T1)	*1*	*1045*	*2*
	Uterine artery Doppler PI (T2)	2	9869	14
	EFW	3	243,130	1721
Poon 2011 predictors + previous PE + previous stillbirth + previous SGA baby	*PlGF*	*1*	*1045*	*2*
	Uterine artery Doppler PI (T2)	*1*	*1045*	*2*
	EFW	**2**	**234,306**	**1709**
Poon 2011 predictors + previous PE + previous stillbirth + EFW	*PlGF*	*1*	*8824*	*12*
	Uterine artery Doppler PI (T2)	*1*	*8824*	*12*

Notes
PI, pulsatility index; PlGF, placental growth factor; T1, first trimester; T2, second trimester; T3, third trimester.
Poon 2011 predictors = gestational age, mother's weight, mother's height, mother's age, parous, smoking, ethnicity (white, black, Asian, Hispanic, mixed or other), chronic hypertension, diabetes, assisted conception.
Bold text = variables carried through to the next iteration; red italics = variables excluded since only one study measured the combination with that variable.

Chapter 5 External validation of existing models

Identification of existing prediction models

We identified 119 prediction models (55 articles) for fetal growth and birthweight (see *Figure 2*). No model reported FGR as pre-specified by us. Of the eleven models that predicted birthweight on a continuous scale, eight (73%) included predictors not reported in the IPPIC cohorts IPD,[59,158,165–167] and two (18%) included combinations of variables not available in the IPPIC IPD cohorts and could not be externally validated.[1,168] One model (Poon 2011) was eligible for external validation using the IPPIC cohorts.[59]

Characteristics of the validated model

The Poon 2011 model predicted birthweight (with a $\log_{10}$ transformation) on a continuous scale and included only clinical characteristics as predictors. The model equation is given below in *Table 4*. The model included gestational age at delivery, mother's weight, height, age, parity, smoking status, ethnicity (white, black, Asian, Hispanic, mixed or other), pre-existing chronic hypertension, diabetes and assisted conception. Gestational age at delivery had the largest impact on predicted birthweight, with an increase in expected birthweight for each week increase in gestational age.

Characteristics of the IPPIC validation cohorts

At database lock for external validation of existing models on 31 January 2020, the IPD of 87 cohorts had been harmonised and were available in the IPPIC data repository. IPD from 10% (9/87) of the cohorts [Allen, ALSPAC (Avon Longitudinal Study of Parents and Children), Baschat, Generation R, Odibo, Rumbold, JSOG (Japan Society of Obstetrics and Gynecology), STORKG, POP][18,75,77,80,87,106,107,120,128] contained all predictor variables and outcomes allowing for external validation of the Poon 2011 prognostic model. Two of the nine cohorts included only nulliparous women.[18,128] The proportion of nulliparous women ranged from 46% to 62% across the remaining cohorts. Five of the included studies were prospective cohorts (Allen, Baschat, Odibo, STORKG, POP),[18,80,87,107,120] three were from prospective registry datasets (ALSPAC, Generation R, JSOG),[75,77,106] and one was a cohort from a randomised trial (Rumbold).[128] All cohorts consisted of unselected pregnant women, except the Rumbold cohort which included only low-risk pregnant women. The median gestational age at delivery was similar across all the cohorts. Most cohorts that recorded ethnicity predominantly consisted of white women, apart from Allen, Baschat and JSOG (47% Asian, 47% black and 100% Eastern Asian included as 'other ethnicity', respectively).

Summary characteristics for the cohorts used in the external validation of the Poon 2011 model are shown in *Table 5*. The greatest proportion of observations with at least one missing value for the variables of interest was observed in ALSPAC (89% incomplete); where mother's height and weight, or birthweight of baby (outcome) were most commonly missing. As the required number of imputations, *m*, was set to at least the proportion of incomplete observations,[60] this informed a minimum requirement of 89 imputed data sets for each study. We chose to impute 100 times for each study, for completeness and to fulfil this requirement. Detailed study characteristics of included IPPIC cohorts used for external validation are provided in *Appendix 1*. Risk of bias assessment of the cohorts by the PROBAST tool considered all cohorts to be at low risk of bias in the domains of participant selection and outcome reporting. All cohorts except the JSOG cohort were considered to be at low risk of bias for the domain

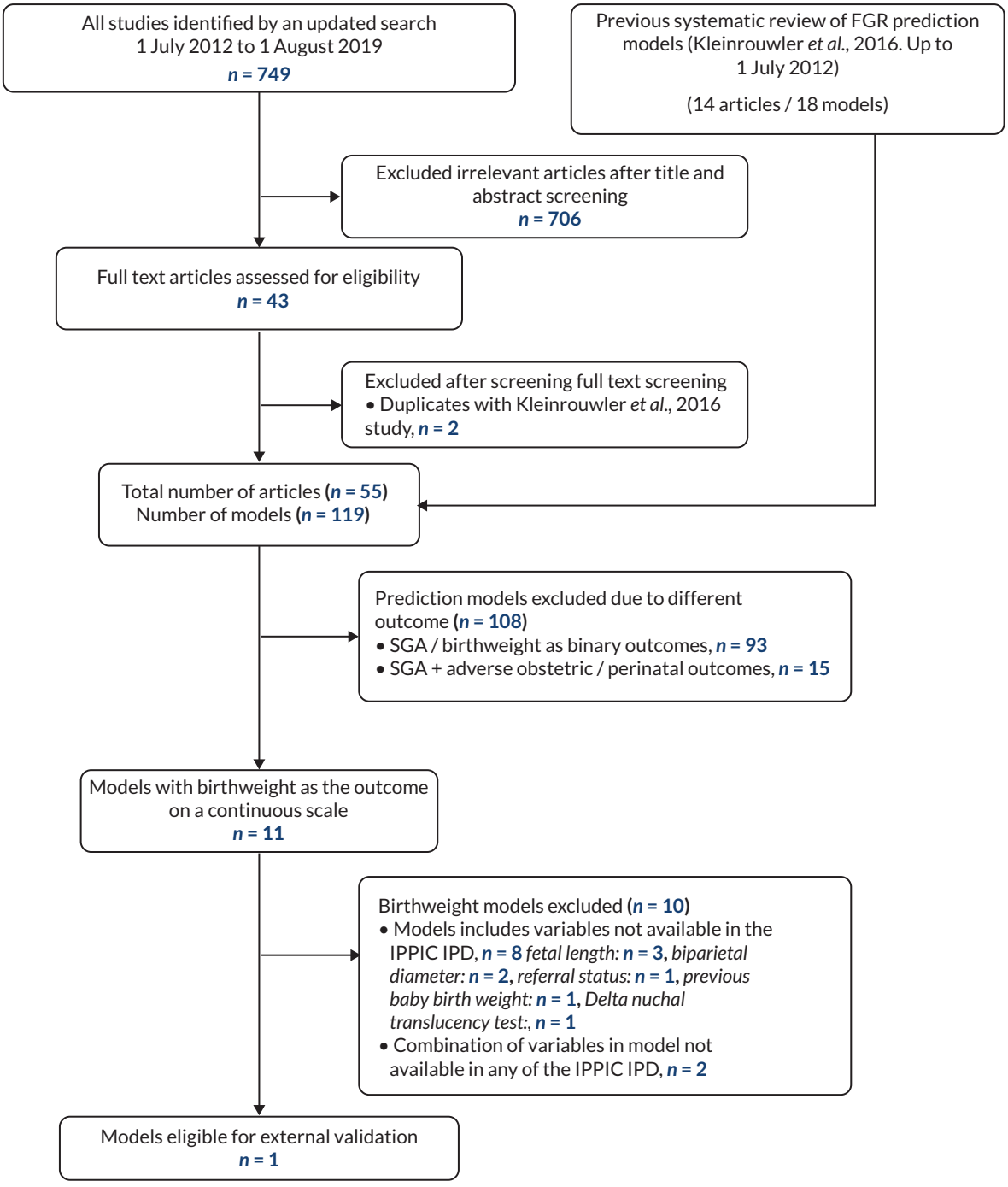

FIGURE 2 Flow chart of identification of eligible FGR prediction models for external validation.

TABLE 4 Poon 2011 model equation[59]

$\log_{10}$ (birthweight) = −0.935219 + 0.186853 (gestational age at delivery, weeks) − 0.002078 (gestational age at delivery, weeks)^2 + 0.003726 (weight, kg) − 0.000030 (weight, kg)^2 + 8.820640e^{-08} (weight, kg)^3 + 0.000965 (height, cm) + 0.001466 (age, years) − 0.000026 (age, years)^2 + 0.016986 (if parous) − 0.024867 (if smoker) − 0.021769 (if African ethnicity) − 0.017824 (if South Asian ethnicity) − 0.005543 (if East Asian ethnicity) − 0.009063 (if mixed ethnicity) − 0.020995 (if chronic hypertension) + 0.03143 (if diabetes) − 0.004015 (if assisted conception)

TABLE 5 Summary characteristics of cohorts used in the external validation of the Poon 2011 model

	Allen[80]	ALSPAC[75]	Baschat[87]	Generation *R*[106]	Odibo[120]	Rumbold[128]	JSOG[77]	STORKG[107]	POP[18]
Number of pregnancies	1045	15,444	1704	8824	1200	1877	406,286	823	4212
Complete (%)	99	11	99	78	95	89	73	46	96
Gestational age at delivery weeks, median (IQR)	40 (39.3–40.6)	40 (39–41)	39.1 (37.9–39.9)	40.1 (39–41)	39.1 (38–39.6)	40 (39–41)	38 (37–40)	40 (38.9–40.9)	40.3 (39.1–41.1)
Weight, kg, median (IQR)	62 (55–69)	55 (50–60)	71.8 (61.3–87.9)	67 (60.5–76)	68.9 (59.9–83.9)	66 (58.5–76)	52 (47–57)	64.55 (56.9–72.9)	66 (59–75)
Height, cm, mean (SD)	161.5 (7.4)	164.3 (6.8)	164 (7)	167.2 (7.4)	164.6 (6.8)	165.3 (6.7)	158.3 (5.5)	163.6 (6.7)	165.2 (6.4)
Age, years, mean (SD)	29.9 (5.1)	27.8 (4.9)	30.2 (6.5)	29.7 (5.3)	31.5 (5.6)	26.4 (5.7)	32.2 (5.4)	29.9 (4.9)	29.9 (5.1)
Nulliparous	461 (44.11)	5828 (37.74)	736 (43.19)	4834 (54.78)	518 (43.17)	0 (0)	210,896 (51.91)	381 (46.29)	0 (0)
Smokers	38 (3.64)	2645 (17.13)	162 (9.51)	1438 (16.3)	97 (8.08)	364 (19.39)	10,952 (2.7)	50 (6.08)	211 (5.01)
Ethnicity									
White	398 (38.09)	12,075 (78.19)	775 (45.48)	4933 (55.9)	735 (61.25)	1777 (94.67)	–	379 (46.05)	3900 (92.59)
Black	108 (10.33)	131 (0.85)	803 (47.12)	2146 (24.32)	325 (27.08)	3 (0.16)	–	62 (7.53)	25 (0.59)
Asian	495 (47.37)	113 (0.73)	88 (5.16)	496 (5.62)	94 (7.83)	1 (0.05)	–	200 (24.3)	91 (2.16)
Hispanic	–	–	27 (1.58)	–	23 (1.92)	1 (0.05)	–	12 (1.46)	–
Mixed	12 (1.15)	–	–	–	22 (1.83)	4 (0.21)	–	–	1 (0.02)
Other	30 (2.87)	82 (0.53)	11 (0.65)	767 (8.69)	1 (0.08)	87 (4.64)	406,286 (100)	170 (20.66)	195 (4.63)
Chronic hypertension	10 (0.96)	1822 (11.8)	162 (9.51)	147 (1.67)	109 (9.08)	9 (0.48)	3421 (0.84)	13 (1.58)	220 (5.22)
Diabetes	11 (1.05)	126 (0.82)	81 (4.75)	33 (0.37)	58 (4.83)	8 (0.43)	2926 (0.72)	–	16 (0.38)
Assisted conception	23 (2.2)	365 (2.36)	35 (2.05)	140 (1.59)	59 (4.92)	49 (2.62)	57,082 (14.05)	13 (1.58)	184 (4.37)
Birthweight, g, mean (SD)	3298.3 (524.5)	3347.7 (608.7)	3147.5 (674.6)	3391.1 (578.4)	3227.9 (676)	3382 (608.9)	2840.4 (581.1)	3418.3 (570.1)	3401 (534.5)

Values are number (%) unless otherwise stated.

of predictor reporting, which had an unclear risk of bias because not enough information was available to make the assessment (see *Appendix 2*).

Performance of existing model in predicting birthweight: external validation and meta-analysis

Average calibration across imputations

Calibration plots (with calibration curves) of the Poon 2011 model were generated in each IPD cohort separately, for each imputation, to assess the similarity of observed and predicted birthweights across the full range of predicted values. A comparison of the observed birthweight distribution and the predicted birthweight distribution in each cohort is given in *Appendix 3*, *Figures 21 and 22*. As calibration plots were very similar on visual inspection across imputations, it was concluded that predictions were similar enough across imputations for pooling to be appropriate. We present in *Figures 3* and *4* calibration plots for the Poon 2011 model in each cohort, comparing the observed birthweight to the average predicted birthweight across imputations. These are presented on the more clinically interpretable birthweight (g) scale (see *Figure 3*), and on the original model scale of $\log_{10}$ (birthweight) (see *Figure 4*), which allows better focus on the birthweights at the lower end of the predicted scale, where pregnancies at higher risk of FGR are more likely to be seen.

On both outcome scales, the light blue LOWESS smoothed calibration curves can be seen to lie close to the diagonal (where expected equals observed outcome value) for all cohorts, suggesting impressive calibration performance on average across individuals from all populations included. We clearly see, though, from the individual points (green) in *Figure 3*, that for predictions at the higher end of the scale (where the bulk of the observations lie), there is a large variation in observed birthweights compared to a relatively narrow range of predictions for all datasets. For example, in the POP[18] cohort predictions in the range of 2500 g to 4000 g correspond to observed birthweights in the range 2000 g to 5000 g. While the model predicts well on average within datasets, there is still some miscalibration in the higher range for some observations.

Calibration plots and curves on the original model scale ($\log_{10}$ birthweight) also show this wider spread of observed values at the upper end of the scale, but this is less pronounced due to the log scale. When focusing on the lower predicted birthweights, those at higher concern regarding FGR, we see more clearly on this scale that calibration in generally good in this clinically important range. Given the clinical requirement of identifying low-birthweight babies at risk of FGR for early intervention, good calibration on average and small variation in predicted birthweights in the lower ranges make the model promising with this use in mind.

Pooled calibration across external validation cohorts

The Poon 2011 model showed reasonable calibration overall in each of the validation cohorts. Individual calibration slopes ranged from 0.91 (95% CI 0.83 to 0.98) in the Allen cohort, to 1.05 (95% CI 1.01 to 1.08) in the POP cohort, suggesting only a small and potentially unimportant miscalibration on average in terms of the slope (as seen visually within the calibration plots and smoothed calibration curves).

The pooled calibration slope across all cohorts of 0.97 (95% CI 0.94 to 1.01, $\tau^2 = 0.0018$) (see *Figure 5*, panel A) implies that the model is well calibrated across cohorts (given the summary calibration slope very close to the ideal value of 1, and its CI also crosses 1). There was also some heterogeneity evident across cohorts; for example, with a 95% prediction interval for the calibration slope in a new study of 0.87 to 1.08, when considering predictions on the birthweight scale. However, this range is still very narrow, and generally miscalibration is predicted to be quite small as measured by the slope.

On the original $\log_{10}$ (grams) scale of the model (see *Figure 5*, panel B), a summary calibration slope of 0.93 (95% CI 0.90 to 0.96, $\tau^2 = 0.0012$) also suggests slight overfitting, with a small amount of

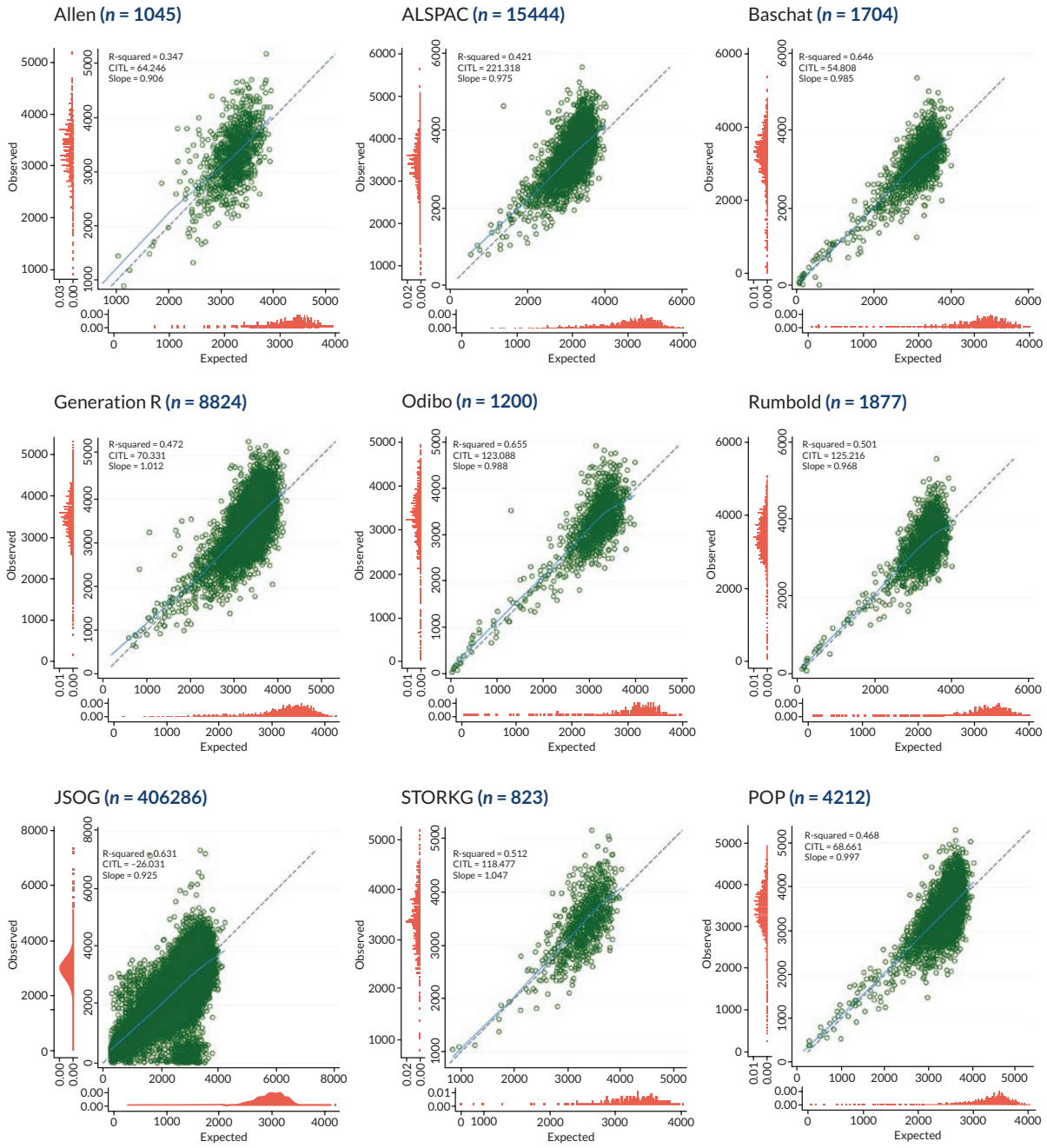

FIGURE 3 Average calibration plots across imputations for individual cohorts on external validation of the Poon 2011 model, with the observed birthweight (g) plotted against predicted birthweight. The dashed line shows perfect calibration (where observed birthweight equals expected birthweight), while the blue line gives the smoothed calibration slope across all pregnancies.

heterogeneity across cohorts. The 95% prediction interval suggests that the calibration slope in a new study would be between 0.84 and 1.02. In practice, predictions of interest are on the grams scale and so we suggest it is better to focus on the previous results.

In the individual cohorts, the smallest CITL value of −26.4 g (−27.5 to −25.3) suggests systematic over-prediction of birthweight on average in the JSOG cohort of 26.4 g, while the largest value suggests an under-prediction of 220.3 g (206.5 to 234.0) on average in the ALSPAC cohort.

The pooled CITL of 90.4 g (37.9 g to 142.9 g, $\tau^2 = 4578$ g^2), when summarised across all gestational ages (*Figure 6*) showed systematic under-prediction of birthweight by around 90.4 g. This is reflected by

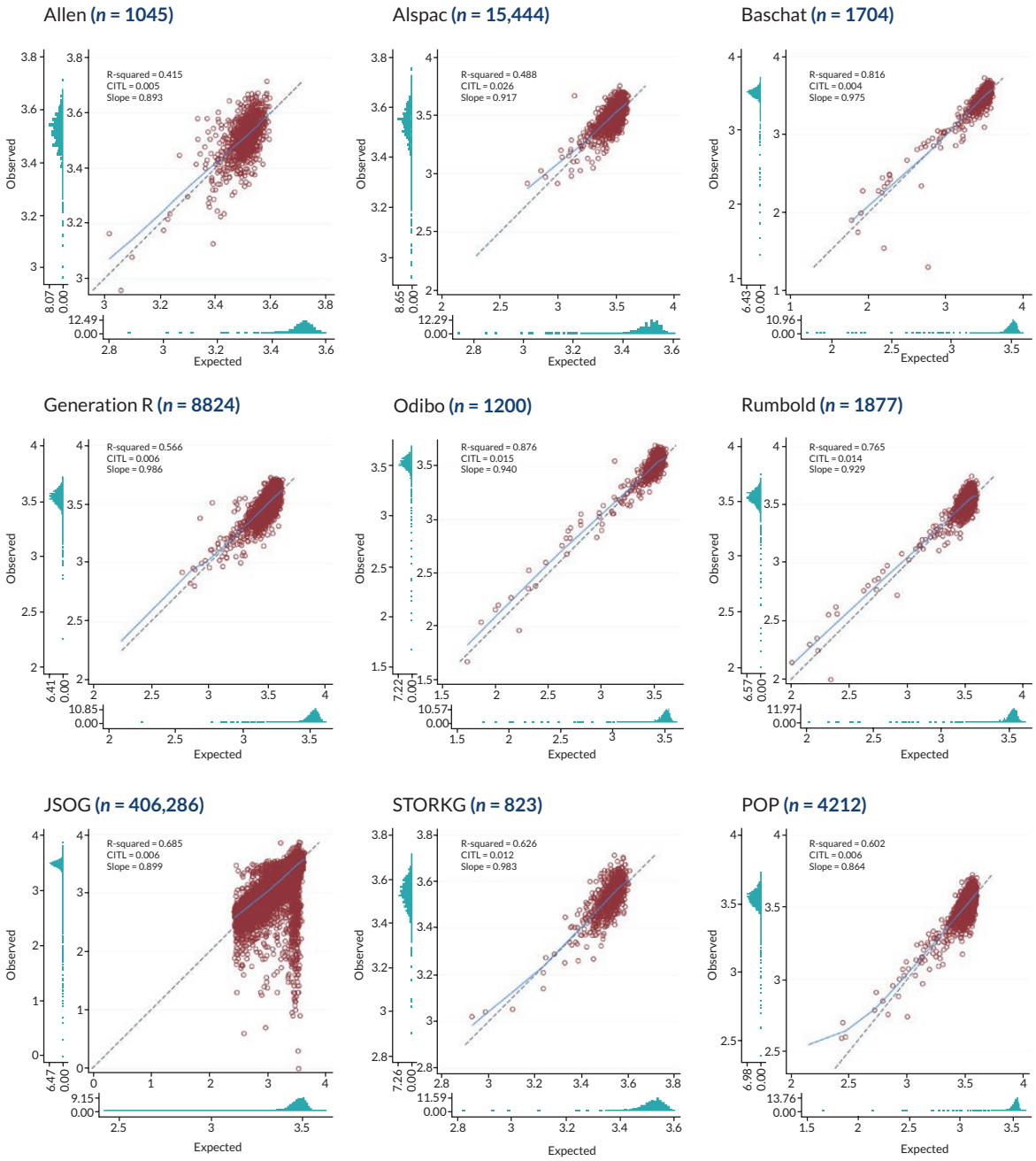

FIGURE 4 Average calibration plots across imputations for individual cohorts on external validation of the Poon 2011 model, with the observed $1og_{10}$ birthweight plotted against predicted $1og_{10}$ birthweight. The dashed line shows perfect calibration (where observed value equals expected value), while the blue line gives the smoothed calibration slope across all pregnancies.

the calibration curves being slightly above the 45° line of perfect calibration in most cohorts. The Poon 2011 model showed moderate between-study heterogeneity in CITL performance, with $\tau^2 = 67.7$ g, and a 95% prediction interval suggesting that we would expect a CITL for a new study to be between −78.4 g and 259.2 g.

Assessing CITL separately by gestational age at delivery (see *Figure 7*) showed that this average under-prediction was consistent across gestational age groups, with the pooled CITL ranging from 94.2 g (95% CI 23.6 g to 164.8 g) in those born 32- and 36-weeks' gestation, to 108.5 g (95% CI −18.5 g to 235.4 g) in those born before 28 weeks. Uncertainty was much higher (with wider CIs for pooled CITL) for estimates at earlier gestational ages for delivery, due to the lower number of observed births before 32 weeks in all cohorts.

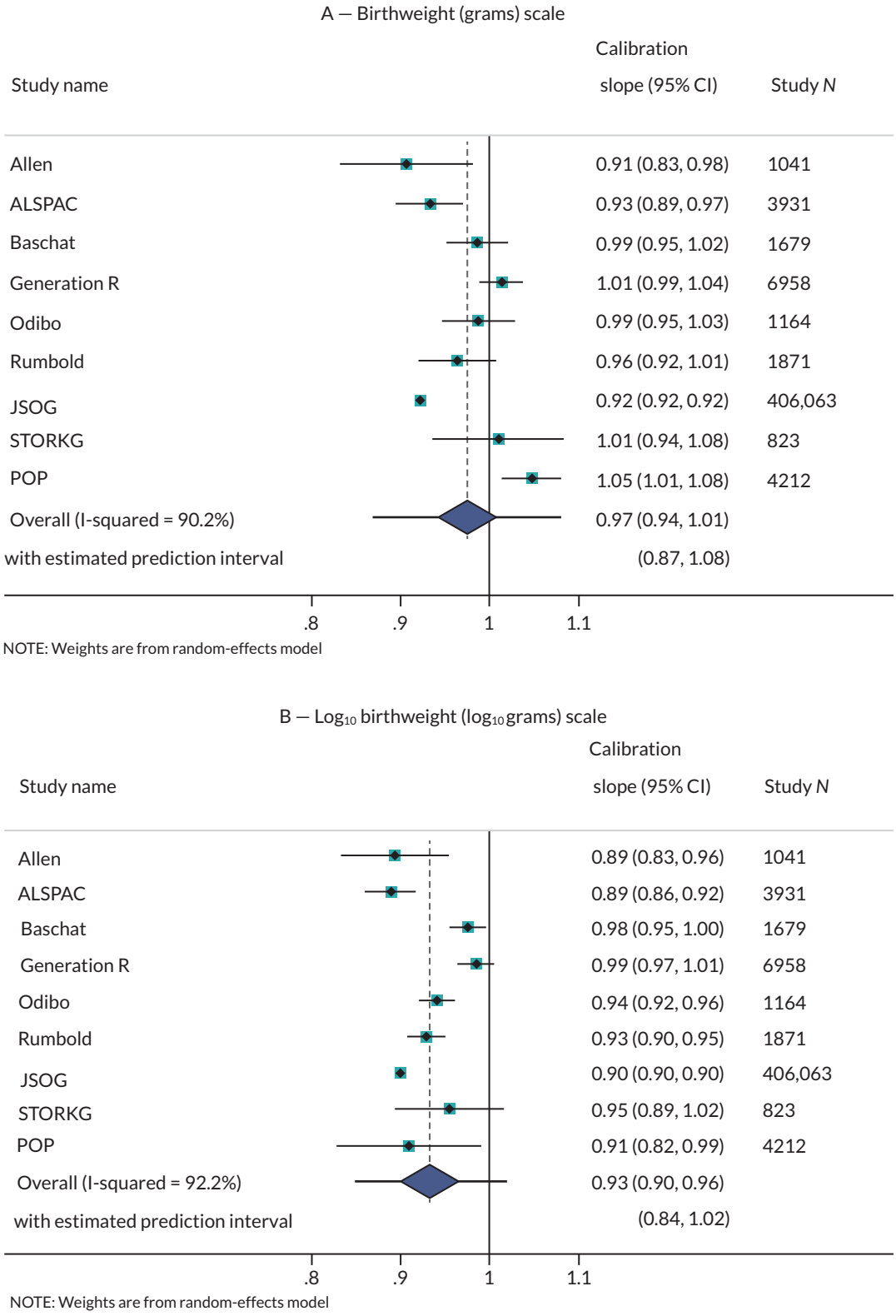

FIGURE 5 Forest plot for the calibration slope of the Poon 2011 model across external validation datasets for predictions made on the birthweight (g) scale (panel A) and the $\log_{10}$ birthweight ($\log_{10}$ grams) scale.

There was moderate to high heterogeneity seen between cohorts in the meta-analysis in all gestational age groups, with relatively wide 95% prediction intervals for all groups. For example, the prediction interval for CITL in those with gestational age <28 weeks at delivery suggests the new study may under-predict birthweight by up to 402.5 g or over-predict birthweight by up to 185.6 g in a new (but similar)

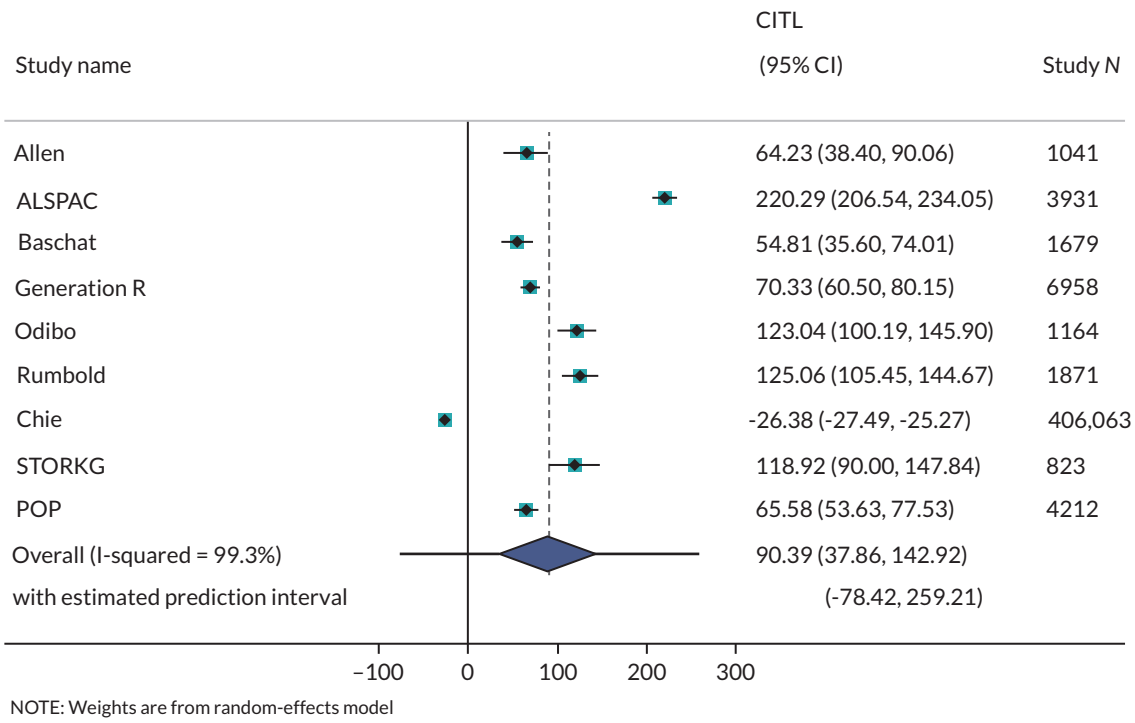

FIGURE 6 Forest plot for the CITL across cohorts (g).

cohort. Given the small average birthweights for babies born at this gestational age, such differences between predicted and observed birthweights are extremely large.

Summary of calibration of the Poon 2011 model

A summary of the meta-analysis results for the calibration slope and the CITL across different gestational age groups is given in *Table 6*. Both calibration measures are important to be considered in combination to assess the calibration performance of a prediction model, and thus a scatter plot including both measures on the individual dataset level is given in *Figure 8*. While no study shows perfect calibration by either measure, the cluster of points in *Figure 8* demonstrates how the Poon 2011 model consistently under-predicts birthweight across cohorts (with the exception of JSOG), regardless of whether the associated calibration slope implied under- or over-fitting. The JSOG dataset can be seen to be an outlier, with one of the lowest calibration slope estimates, and was the only cohort to suggest an over-prediction of birthweight on average when using the Poon 2011 model.

On average across external validation cohorts, the calibration slope of the Poon 2011 model was impressive when including all gestational age groups in the analysis, suggesting minimal overfitting of the model on average (pooled calibration slope: 0.97) across all age groups. Most overfitting was seen for those with gestational age 28–31 weeks, where a pooled calibration slope of 0.89 suggests that predictions were too extreme.

Calibration-in-the-large was also promising on average, with an average under-prediction of birthweight by 90.4 g (where under-prediction is clinically preferable in the determination of FGR risk). This average underprediction was consistent across gestational ages, which would have more of a relative impact on the usefulness of predictions for smaller babies born at earlier gestational ages.

Calibration curves for the Poon 2011 model reflect the similarity of observed and predicted birthweights suggested from the promising calibration slope and CITL values. The LOWESS smoothed calibration curves can be seen to lie close to the diagonal (where expected equals observed outcome value) for all cohorts, suggesting impressive calibration performance on average across individuals from all populations included.

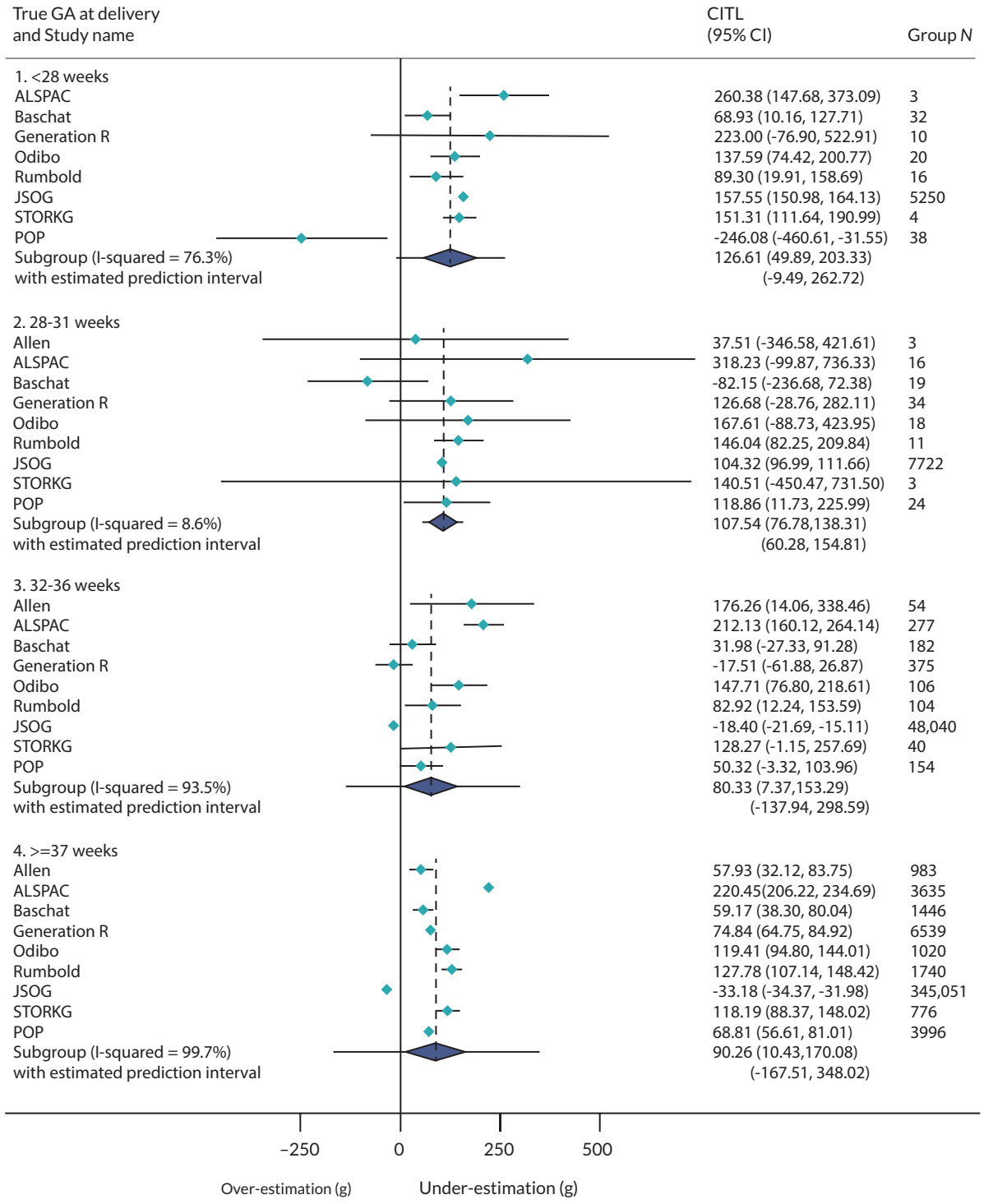

FIGURE 7 Forest plot for CITL across cohorts, grouped by gestational age at delivery. GA, gestational age.

Summary

In summary, from 119 prediction models for fetal growth and birthweight identified in our literature search, no prediction models were found to predict the probability of our predefined definition of FGR. One birthweight model could be externally validated. The Poon 2011 model predicts $\log_{10}$ (birthweight) using 10 variables based on maternal characteristics only.

TABLE 6 Pooled calibration measures by gestational age at delivery

Gestational age at delivery	Number of datasets in meta-analysis	Performance measure	Pooled estimate	CI	Prediction interval	τ^2
Any	9	Calibration slope	0.974	0.938 to 1.011	0.868 to 1.081	0.0018
		CITL	90.39 g	37.9 to 142.9	−78.4 to 259.2	4578
<28 weeks	8	Calibration slope	1.163	0.893 to 1.432	0.53 to 1.79	0.0531
		CITL	126.61 g	49.9 to 203.3	−9.5 to 262.7	2041
28–31 weeks	9	Calibration slope	0.894	0.850 to 0.937	0.85 to 0.94	0.0000
		CITL	107.54 g	76.8 to 138.3	60.3 to 154.8	222
32–36 weeks	9	Calibration slope	1.043	0.887 to 1.199	0.62 to 1.47	0.0276
		CITL	80.33 g	7.4 to 153.3	−137.9 to 298.6	7519
≥37 weeks	9	Calibration slope	0.907	0.838 to 0.976	0.70 to 1.11	0.0067
		CITL	90.26 g	10.4 to 170.1	−167.5 to 348.0	10,685

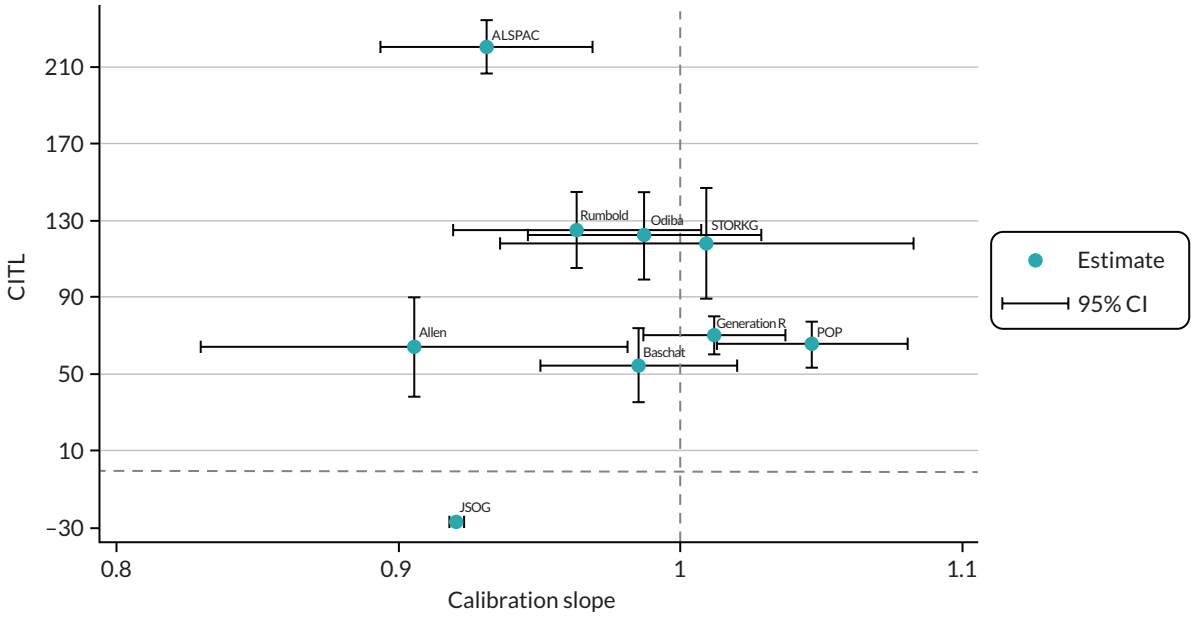

FIGURE 8 Scatterplot comparing CITL and the calibration slope of the Poon 2011 model, as estimated in each cohort. The dotted lines indicate perfect calibration by each measure.

External validation of the Poon 2011 model was possible in 9 cohorts from the IPPIC repository, containing data on 441,415 pregnancies. Calibration of the Poon 2011 model was promising, with the pooled calibration slope only slightly lower than one on average across cohorts. However, there was some heterogeneity in the calibration performance across cohorts, with the calibration slope in individual cohorts lying slightly above or below the ideal value of one (implying predictions are slightly too extreme in some cohorts, and not quite extreme enough in others).

The model predictions could also be systematically too low or too high depending on the cohorts used to validate the model, although the Poon 2011 model was most seen to slightly under-predict birthweight. Under-prediction was by around 100 g on average across datasets, regardless of gestational

age at delivery. The relative effect of this under-prediction would be greater in babies born at younger gestational ages, where expected birthweight is lower.

However, calibration was very good in general. Hence, due to the reasonably good performance of the Poon 2011 model on average across cohorts, we concluded that it would be illogical to begin building a new prediction model from scratch. Therefore, in the next chapter, we update the Poon 2011 model for predicting birthweight by using their included predictor variables as a basis for an updated model predicting the probability of FGR in pregnant women. By considering additional variables, agreed by clinical consensus, we further develop a model for predicting birthweight to ascertain whether the inclusion of new variables might improve the consistency of calibration across populations.

Chapter 6 Development and validation of fetal growth restriction and birthweight models

In this chapter we discuss the results of the development and validation of two new models to predict (1) FGR; and (2) birthweight, using the IPPIC datasets. The full methods for the development and validation of these models are included in *Chapter 3*, *Recalibration of existing fetal growth restriction prediction models*.

Characteristics of IPPIC cohorts included in the IPD meta-analysis

At database lock for the development of the FGR and birthweight models on 31 August 2020, 94 cohorts were available in the IPPIC data repository. After prioritisation of predictors from existing literature and clinical consensus (see *Prioritisation of candidate predictors of fetal growth restriction: Delphi survey findings*), IPD from four cohorts were selected as giving the best combination of predictor variables while maximising the numbers of cohorts, participants, and events for model development (see *Prioritisation of candidate predictors of fetal growth restriction: Delphi survey findings*). Three of the included cohorts were from prospective observational studies [Allen, STORKG, NICHD CSL (National Institute of Child Health and Human Development Consortium on Safe Labour)][80,107,164] and included unselected pregnant women. The Rumbold cohort was from a randomised trial and included low-risk women.[128]

One cohort included only nulliparous women,[128] while the remaining three had proportions of nulliparous women ranging from 40% to 56%. Across cohorts, the most common ethnicity was white (50%), followed by black (22%). Hispanic mothers were also well represented (17%) due to the high proportion of this ethnicity in the NICHD CSL cohort. The median gestational age of delivery was similar across all the cohorts (39–40 weeks), as well as the mean birthweight. The mean birthweight for all cohorts lay within a range of around 200 g, from 3199.8 g in NICHD CSL, up to 3418.3 g in STORKG. The composite FGR outcome was rare in all cohorts: notably only two pregnancies (0.2%) in the Allen cohorts and no women in the STORKG cohort met our criteria for FGR with complications. Across all four cohorts, 1729 (0.7%) pregnancies reported the outcome of FGR with complications, of these 1389 (80.3%) delivered before 32 weeks, 505 (29.2%) were stillbirths and 420 (26.7%) resulted in a neonatal death.

Detailed study characteristics of IPPIC cohorts used in model development are provided in *Appendix 1*, risk of bias assessment of the cohorts using the PROBAST tool is provided in *Appendix 2* and plots of predictor distributions across the model development cohorts are provided in *Appendix 4*, *Figures 23–28*.

Missingness and multiple imputation

The birthweight outcome was rarely missing across cohorts, with the maximum proportion missing seen in the STORKG cohort at just 4.6%.[107] The composite FGR outcome was based upon the gestational age at delivery and birthweight (both of which were mostly complete in all cohorts), and complications of preterm birth (defined by gestational age at delivery, mostly complete), stillbirth (complete in all cohorts), or neonatal death. Neonatal death was well recorded in two of the cohorts (Rumbold, NICHD CSL),[128,164] but was entirely missing in the remaining two (Allen, STORKG).[80,107] Given the rarity of neonatal death in the underlying populations (0.4%) and the small size of these datasets, we chose to assume neonatal death was not observed for all pregnancies included in these two datasets. Due to the rarity of neonatal death in combination with birthweight <10th centile, we would not expect this assumption to greatly influence the model estimates.

Summary characteristics for the cohorts used in development of the FGR and birthweight models, including proportions missing for each predictor, are shown in *Table 7*. The greatest proportion

TABLE 7 Characteristics of cohorts included in prediction model development

	Allen[80]		Rumbold[128]		STORKG[107]		NICHD CSL[164]		Total	
		Missing		Missing		Missing		Missing		Missing
N	1045	13 (1.2)	1877	196 (10.4)	823	442 (53.7)	233,483	222,845 (95.4)	237,228	223,496 (94.2)
Gestational age at delivery (weeks), median (IQR)	40 (39.3–40.6)	1 (0.1)	40 (39–41)	–	40 (38.9–40.9)	22 (2.7)	39 (38–40)	7929 (3.4)	39 (38–40)	7952 (3.4)
Mother's weight, kg, median (IQR)	62 (55–69)	5 (0.5)	66 (58.5–76)	103 (5.5)	64.6 (56.9–72.9)	421 (51.2)	66.7 (57.6–80.3)	31,314 (13.4)	66.7 (57.6–80)	31,843 (13.4)
Mother's height, cm, mean (SD)	161.5 (13.3)	–	165.3 (15.7)	138 (7.4)	163.6 (13.3)	–	163.3 (6.6)	37,567 (16.1)	163.3 (8)	37,705 (15.9)
Mother's age, years, mean (SD)	29.9 (7.4)	1 (0.1)	26.4 (6.7)	–	29.9 (6.7)	–	27.7 (7.4)	339 (0.1)	27.7 (7.4)	340 (0.1)
Nulliparous	584 (55.9)	–	1877 (100)	–	381 (46.3)	–	93,545 (40.1)	0 (0)	96,387 (40.6)	0 (0)
Smoked during pregnancy	38 (3.6)	–	364 (19.4)	39 (2.1)	50 (6.1)	–	15,547 (6.7)	0 (0)	15,999 (6.7)	39 (0)
Ethnicity		2 (0.2)		4 (0.2)		–		9557 (4.1)		9563 (4.0)
White	398 (38.1)		1777 (94.7)		379 (46.1)		116,000 (49.7)		118,554 (50)	
Black	108 (10.3)		3 (0.2)		62 (7.5)		52,518 (22.5)		52,691 (22.2)	
Asian	495 (47.4)		1 (0.1)		200 (24.3)		9487 (4.1)		10,183 (4.3)	
Hispanic	–		1 (0.1)		12 (1.5)		40,409 (17.3)		40,422 (17)	
Mixed	12 (1.1)		4 (0.2)		–		347 (0.1)		363 (0.2)	
Other	30 (2.9)		87 (4.6)		170 (20.7)		5165 (2.2)		5452 (2.3)	
History of hypertension	10 (1)	–	9 (0.5)	–	13 (1.6)	–	4589 (2)	0 (0)	4621 (1.9)	0 (0)

TABLE 7 Characteristics of cohorts included in prediction model development (*continued*)

	Allen[80]		Rumbold[128]		STORKG[107]		NICHD CSL[164]		Total	
		Missing		Missing		Missing		Missing		Missing
History of diabetes	11 (1.1)	–	8 (0.4)	–	–	–	4946 (2.1)	7878 (3.4)	4965 (2.1)	7878 (3.3)
Assisted conception	23 (2.2)	–	50 (2.7)	39 (2.1)	13 (1.58)	–	1472 (0.6)	109,799 (47)	3354 (1.4)	109,838 (46.3)
Any previous PE	17 (1.6)	–	–	–	–	–	10,131 (4.3)	31,545 (13.5)	10,148 (4.3)	31,545 (13.3)
Any previous stillbirth	12 (1.1)	–	–	–	8 (1)	–	2029 (0.9)	96,159 (41.2)	2049 (0.9)	96,159 (40.5)
Any previous SGA baby	67 (6.4)	–	–	–	31 (3.8)	–	2857 (1.2)	9640 (4.1)	2955 (1.2)	9640 (4.1)
Birthweight (g), mean (SD)	3298.3 (524.5)	4 (0.4)	3382 (608.9)	6 (0.3)	3418.3 (570.1)	38 (4.6)	3199.8 (644.1)	2674 (1.1)	3202 (643.4)	2722 (1.1)
FGR outcome[a]	2 (0.2)		18 (1)		0 (0.0)		1709 (0.7)		1729 (0.7)	
Preterm birth (<32 weeks)	1 (50.0)	–	11 (61.1)	–	0 (0.0)	–	1377 (80.6)	–	1389 (80.3)	–
Stillbirth	1 (50.0)	–	9 (50.0)	–	0 (0.0)	–	495 (29.0)	–	505 (29.2)	–
Neonatal death	–	1045 (100)	4 (23.5)	1 (0.1)	–	823 (100)	416 (26.7)	23,873 (10.3)	420 (26.7)	25,700 (11.0)

a Totals exceed 100% as the components of the FGR outcome are not mutually exclusive.

Note
Values are number (%), unless otherwise stated.

of observations with at least one item of missing information was observed in NICHD CSL (95% incomplete), where conception mode or previous stillbirth were most commonly missing. As the required number of imputations, *m*, was set to at least the proportion of incomplete observations,[60] this informed a minimum requirement of 95 imputed data sets for each study. We again chose to impute 100 times for each study to fulfil this requirement. Details of imputation checks are included in *Appendix 5*, *Figures 29–33*.

Identification of non-linear associations in complete case data

As discussed in *Missing data*, we performed a complete-case analysis to identify potential non-linear associations between continuous predictors and the outcomes. Best-fitting fractional polynomial transformations were assessed in the presence of all other model predictors.

Upon visual inspection of the selected non-linear functions selected for mother's height and mother's weight, there was no improvement in functional fit with FP2 compared to FP1. Therefore, FP1 transformations of $height^3$ and $weight^{-1}$ were taken forward to imputation.

While the best fit for gestational age at delivery (GA) was linear at FP1, FP2 analysis suggested a transformation, in the presence of other predictors, with powers of GA^{-2} and GA^{-2} ln *GA*. The selected FP2 function included a flattening of the curve at the extremes, as can be seen in *Figure 9*, avoiding the negative expected birthweights at lower gestational ages that would arise if assuming a linear association. Instead, the selected FP2 function suggests that birthweights decrease with increasing gestational ages up to about 23 weeks, which is also illogical. However, gestational ages affected by this would be outside the range of the expected model use, and so predictions were unlikely to be affected in practice. While a FP1 transformation (linear fit in this case) might be preferable for the sake of parsimony, we elected

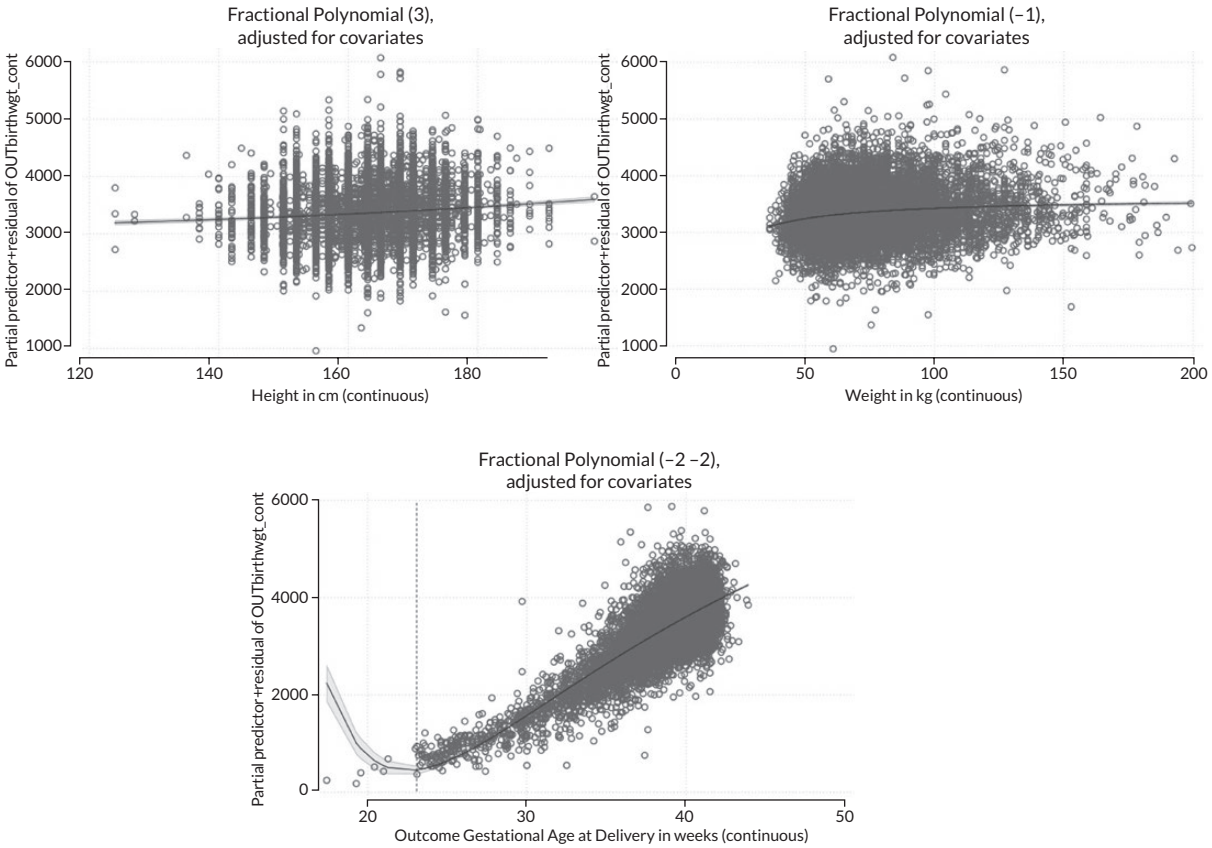

FIGURE 9 Best-fitting fractional polynomial transformations for continuous predictors in complete case data: mother's height (cm), mother's weight (kg) and gestational age at delivery (weeks). Note: Given the shape of selected FP2 transformation, predictions are not clinically relevant when based on assumed gestational ages below 23 weeks (dotted line).

to take forward the possibility of non-linear transformation into the final analysis, to align with the FP2 transformation of gestational age at delivery in the Poon model, which we aimed to update.

Linear functions were selected as the best fit for mother's age compared to both FP1 and FP2 functions, therefore, only the linear term was taken forward. Although the 'best-fitting' functional form was selected for each continuous predictor, it is evident from visual inspection that individual values for each predictor can vary a great deal from the line showing the selected function.

Identification of non-linear associations in multiply imputed data

Imputed datasets were generated including both the linear and the selected non-linear transformation terms for the above continuous variables (mother's height, weight and gestational age at delivery) in the imputation model. Backwards stepwise variable selection procedures were run, comparing models with and without variable transformations at each iteration.

Following variable selection procedures, the imputation was rerun with only the final included transformation terms, to reduce any noise in the imputation model arising from spuriously identified variable transformations. Excluded variables were included in the imputation model on their original scale rather than using any transformation.

The number in brackets shows the best-fitting transformation, where (3) is FP1 with a power of x^3, (−1) is FP1 with the reciprocal (power x^{-1}) and (−2 −2) is FP2 with the powers of x^{-2} and $x^{-2}\ln x$.

Predicting fetal growth restriction IPPIC-FGR prediction model

We developed the IPPIC-FGR model to predict FGR (a binary outcome) using data from all four IPPIC cohorts used to develop the IPPIC-birthweight model. A summary of the predictors retained in the FGR model after variable selection is given in *Table 8*. Spontaneous conception, a history of diabetes and mother's weight were not retained in the model to predict FGR. The weight variable was excluded, with neither the linear nor the transformed terms being below the significance threshold for retention in the model, when considered along with the other model variables.

Gestational age at delivery was retained in the model, allowing predicted FGR risk to be generated conditional on any assumed (clinically relevant) value for gestational age at delivery, or indeed across a range of assumed values, as desired. Conditional on the other model variables, increased gestational age at delivery was associated with a reduced risk of FGR, as were increased mother's height and being of 'other' ethnicity. Being nulliparous, smoking during pregnancy, an increase in mother's age, a history of hypertension, previous PE, previous stillbirth or having had a previous SGA baby all increased FGR risk. All ethnicities other than white or 'other' were also associated with an increased risk of an FGR pregnancy.

An estimate of heuristic shrinkage was calculated in each imputation, and when averaged across imputations was 0.9985, implying very little overfitting in the development data due to the large effective sample size. Given this, it was concluded that application of shrinkage was unnecessary, and so no shrinkage (or associated re-estimation of the intercept term) was applied to the model.

Apparent overall performance and by cohorts

The apparent performance of the model was calculated by applying the FGR model directly back into each dataset using the average intercept term, using the observed gestational age at delivery for each participant (which would be unknown at the time of prediction in practice). On visual inspection, separation of the LP between events and non-events (discrimination) looked promising across datasets (see *Figure 10*). In particular, separation was good between the two LP distribution curves in the NICHD CSL dataset, where the bulk of the model development data originates. Note that no FGR events occurred in the STORKG dataset, hence the absence of the red FGR distribution, and there were only

TABLE 8 Prediction model for FGR with study specific and average intercept terms: model coefficients and odds ratios (OR) with 95% CIs

	Coefficient	OR (95% CI)
Gestational age at delivery (weeks)		
wks^{-2}	−56,010.23	–
$wks^{-2} * \ln wks$	21,652.92	–
Mother's age (years)	0.0104503	1.011 (1.002 to 1.020)
Mother's height (cm)		
cm^3	-1.08×10^{-07}	–
Nulliparous	0.3584681	1.431 (1.265 to 1.619)
Smoker	0.2928371	1.340 (1.116 to 1.609)
Ethnicity		
White	ref	ref
Black	0.4317056	1.540 (1.341 to 1.768)
Asian	0.1813291	1.199 (0.842 to 1.707)
Hispanic	0.2961263	1.345 (1.132 to 1.597)
Mixed	0.9533642	2.594 (0.716 to 9.397)
Other	0.0034091	1.003 (0.713 to 1.412)
History of hypertension	0.3133796	1.368 (1.036 to 1.807)
Any previous PE	0.8867762	2.427 (2.065 to 2.854)
Any previous stillbirth	0.4355474	1.546 (1.066 to 2.241)
Any previous SGA baby	2.16594	8.723 (7.188 to 10.585)
Intercept		
Average	−22.8107	–
NICHR-CSL	−22.8165	–
Allen	−23.4148	–
Rumbold	−21.5836	–
STORKG	−22.8107	–
Heuristic shrinkage	0.9985	

two FGR events in the Allen dataset. In both cases, the distribution of the LPs for pregnancies without FGR is similar to the corresponding distributions from the other two datasets.

Study-specific model performance was not assessed for STORKG, as no FGR events were observed in the dataset: indeed, this study will have had very little weight towards the predictor effect estimates. The apparent predictive performance was pooled across datasets and is reported in *Table 9* along with performance measures calculated in the full dataset, where predictions were calculated using study-specific intercepts. IECV was not used as planned for this binary model due to the low number of FGR outcomes in some of the smaller datasets.

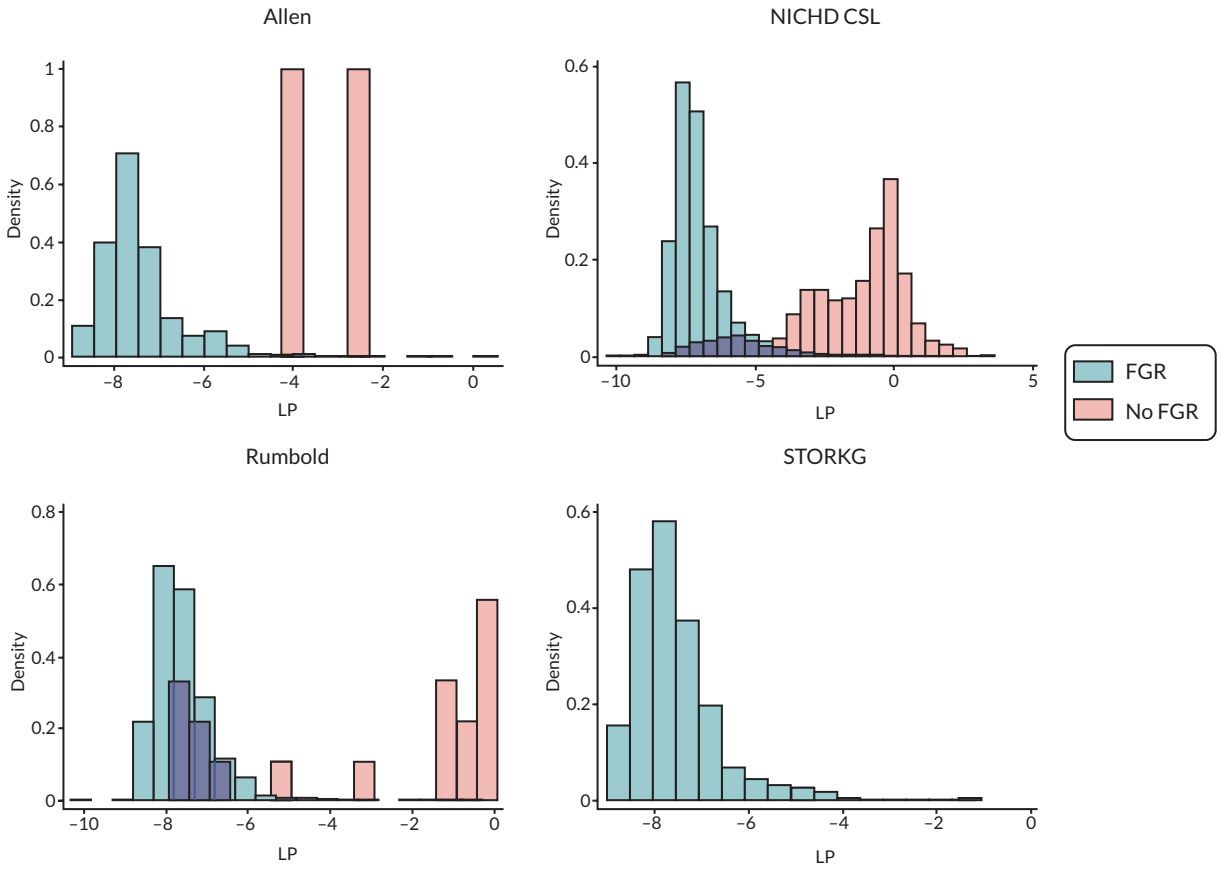

FIGURE 10 Distributions of LP values in the four model development datasets, separated by observed outcome status.

When including all participants (and thus any gestational age at delivery), the pooled *c*-statistic across datasets suggests excellent discrimination, at 0.962 (95% CI 0.508 to 0.998). The low number of events in the Allen dataset gives a misleadingly high estimate of the *c*-statistic, with narrow CIs (given the CI width is dependent on the *c*-statistic value). Apparent discrimination performance was impressive in both Rumbold and NICHD CSL datasets, with *c*-statistic estimates of 0.874 (95% CI 0.737 to 0.945) and 0.962 (95% CI 0.956 to 0.968), respectively.

The calibration slope was also promising, with a pooled apparent performance across datasets of 0.945 (95% CI 0.665 to 1.230) although again with wide CIs, given the small number of studies included in the meta-analysis.

Being the largest of the model development datasets, so contributing most to the number of events in the pooled data, we expect model performance to be at its best for NICHD CSL. Indeed, the model was calibrated best in this dataset by all performance measures, for example with an Observed to Expected ratio of 0.996 (95% CI 0.996 to 0.996) suggesting that the model is well calibrated for predicting FGR in this population, as anticipated.

Calibration plots in *Figure 11* show the apparent calibration performance of the FGR model when applied to all participants in each dataset. Observed and expected FGR proportions are for risk groups by predicted FGR risk. On visual inspection of the smoothed calibration curve over all observations within a dataset, the model appears to overpredict FGR over the full range of predicted probabilities for both the Allen and Rumbold datasets, however, there were only 2 and 18 FGR outcomes, respectively in these datasets.

TABLE 9 Apparent predictive performance measures for the FGR model (applying predictions using the average intercept) for each dataset including all participants regardless of gestational age at delivery) and with pooled effect estimates across datasets

	Pooled estimate	Allen	Rumbold	NICHD CSL	Full data
N (events)	**236,405 (1729)**	**1045 (2)**	**1877 (18)**	**233,483 (1709)**	**237,228 (1729)**
Calibration					
Calibration slope					
Point estimate	0.947	0.829	0.850	1.003	1.000
CI	0.665 to 1.230	0.361 to 1.298	0.669 to 1.031	0.979 to 1.027	0.976 to 1.024
τ^2, 95% CI	0.007 (0.000 to 0.263)	–	–	–	–
CITL					
Point estimate	0.323	−0.604	1.227	−0.006	−0.0001
CI	−1.881 to 2.527	−2.173 to 0.965	0.616 to 1.838	−0.062 to 0.050	−0.056 to 0.056
τ^2, 95% CI	0.612 (0.055 to 14.468)	–	–	–	–
Observed/expected					
Point estimate	0.323	0.634	2.190	0.996	1.000
CI	−1.881 to 2.527	0.633 to 0.636	2.181 to 2.199	0.996 to 0.996	1.000 to 1.000
τ^2, 95% CI	0.393 (0.089 to 6.889)	–	–	–	–
Discrimination					
c-statistic					
Point estimate	0.962	0.990	0.874	0.962	0.961
CI	0.508 to 0.998	0.969 to 0.997	0.737 to 0.945	0.956 to 0.968	0.955 to 0.967
τ^2 (95% CI)	1.373 (0.095 to 28.795)	–	–	–	–
Nagelkerke pseudo R^2					
Median (%)	40.1	30.4	40.1	50.2	50.1
Range	30.4–50.2	30.4–30.5	39.9–40.3	49.9–50.6	49.7–50.4
IQR	30.4–50.2	30.4–30.5	40.0–40.1	50.2–50.3	50.0–50.1

Notes
Prediction intervals not calculated due to small number of cohorts.
Note the full data calculations also include non-event pregnancies from the STORKG cohort.

The model appears to be well calibrated in the NICHD CSL cohort for predicted probabilities of FGR below 0.5. This is in line with the O/E ratio and calibration slope estimates in this study. A similar calibration pattern is seen in the full data, as expected as the NICHD CSL study dominates the total dataset.

Model performance by assumed gestational age at delivery

Given gestational age at delivery is unknown at the time of prediction, clinically relevant timepoints must be chosen prior to prediction generation. This allows predictions of FGR to be produced conditional on various possible delivery times, and thus allows for calculation of distinct probabilities of FGR for a pregnancy at every possible gestational age at delivery.

Discrimination measures presented in *Apparent overall performance and by cohorts* are likely to be optimistic (e.g. *c*-statistic too high), as validation was conducted using predictions generated for a

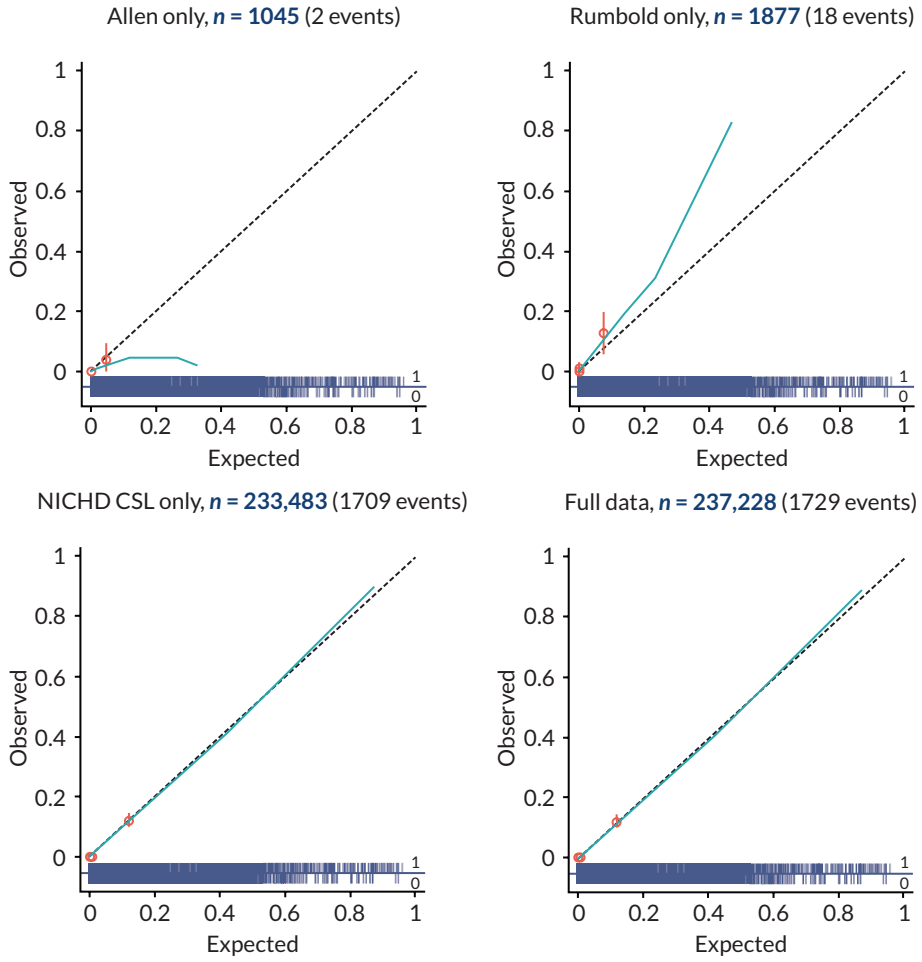

FIGURE 11 Calibration plots of FGR prediction model, in all cohorts combined and in each of the model development datasets individually (apparent calibration performance), based on all participants (regardless of gestational age at delivery). The dashed line shows perfect calibration (where observed proportion equals expected proportion), while the blue line gives the smoothed calibration slope across all pregnancies. Number of pregnancies (events) is given for each. Average predicted risk is shown in 10 groups by predicted risk (green) and smoothed over all individuals (blue).

participant's known delivery time. To give a more complete picture, we further examined the FGR model's predictive performance for scenarios where all predictions in the data were generated conditional on the same assumed gestational age at delivery. Apparent calibration curves are presented in *Figure 12* for the model's calibration performance where all predictions were generated using the same assumed gestational age at delivery of (1) 34 weeks; (2) 36 weeks; (3) 38 weeks; and (4) 40 weeks, for everyone in the data, and then validated against (1) all women (regardless of their gestational age of delivery); and (2) the subset of individuals that actually had those gestational ages of delivery.

Predictions at particular gestational ages and compared to observed FGR status regardless of gestational age at delivery

Predicted probabilities of FGR show greater spread when predictions are made conditional on a gestational age at delivery of 34 weeks, compared to later times. Calibration of the FGR model was best on average when assessed using an assumed gestational age at delivery of 34 weeks for all participants. When generating predictions conditional on the same fixed gestational age for all, the ordering of predicted probabilities did not vary when the assumed gestational age was changed, thus the discriminative ability of the model was consistent across all gestational age values when including validation in all women. When analysed in the largest of the datasets (NICHD CSL), a *c*-statistic of 0.742 (95% CI: 0.729 to 0.754) suggests good discrimination, even in the absence of a known

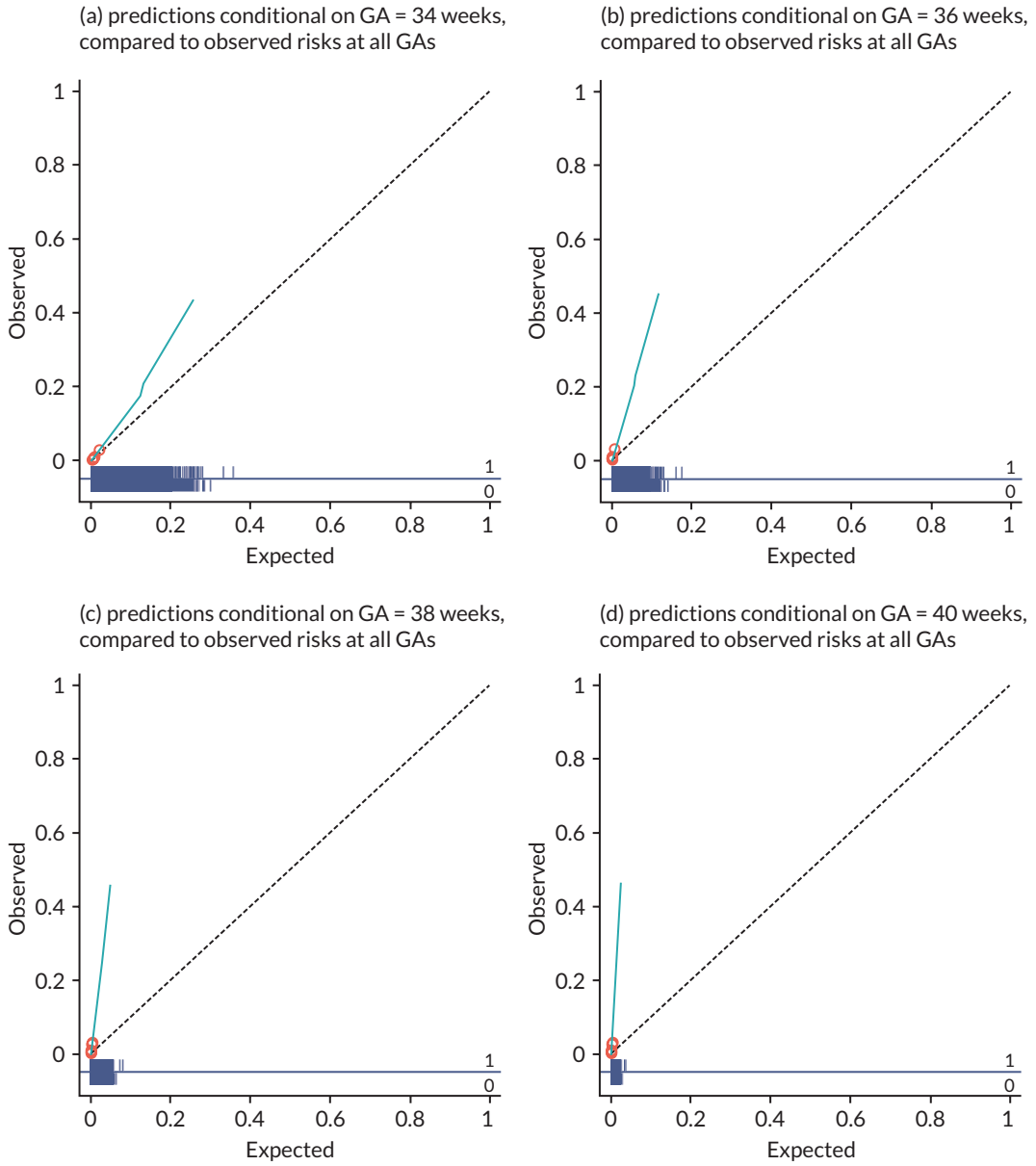

FIGURE 12 Calibration plots of FGR prediction model in all cohorts combined, with predictions generated at the same assumed GA at delivery for every participant, but compared to observed risks at all Gas. Plots are given for assumed GA at delivery of 34 weeks (panel A), 36 weeks (panel B), 38 weeks (panel C), and 40 weeks (panel D), and evaluated against observed FGR status (regardless of gestational age of delivery). The dashed line shows perfect calibration (where observed proportion equals expected proportion), while the blue line gives the smoothed calibration slope across all pregnancies. Average predicted risk is shown in 10 groups by predicted risk (green) and smoothed over all individuals (blue).

gestational age at delivery, when evaluated across all women regardless of their gestational age at delivery. The pooled *c*-statistic across all datasets further suggests good discrimination, at 0.658 (95% CI: 0.262 to 0.913).

Predictions at particular gestational ages and compared to observed FGR status in the subset of participants who actually had that gestational age at delivery

Predictions conditional on an assumed gestational age at delivery were further assessed for calibration performance in the subgroup of participants who truly gave birth at (or close to) that assumed gestational age. The FGR model (generating predictions at a fixed gestational age) was best calibrated in pregnancies with a true gestational age <32 weeks, with calibration curves very close to the diagonal

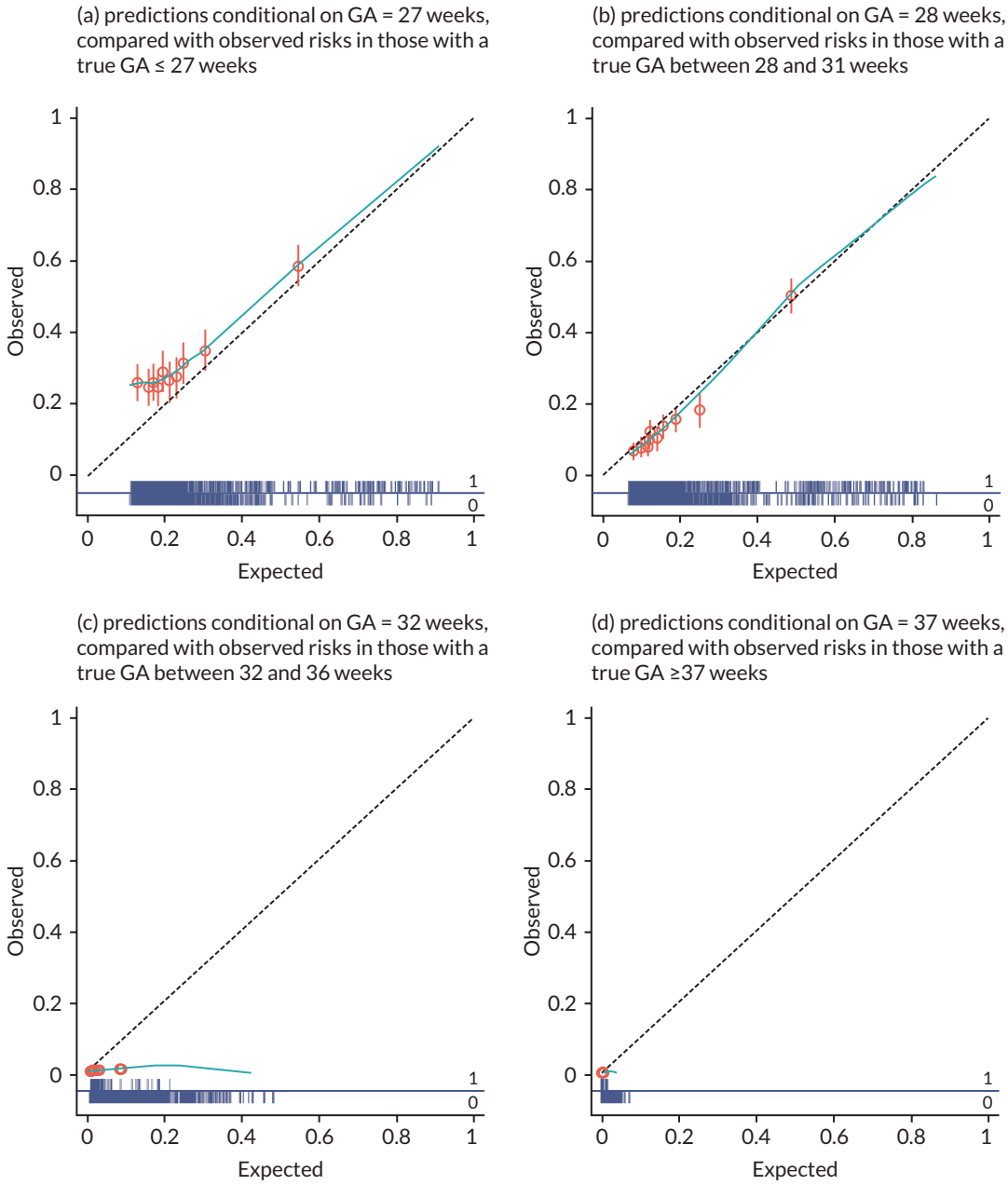

FIGURE 13 Calibration plots of FGR prediction model in subgroups by gestational age at delivery, with predictions generated at the same assumed GA at delivery for every participant and evaluated against observed FGR status in subgroups defined by those with similar (but not identical) actual gestational ages. Plots are given for assumed GA at delivery of 27 weeks in those with a true GA ≤ 27 weeks (panel A), 28 weeks in those with a true GA between 28 and 31 weeks (panel B), 32 weeks in those with a true GA between 32 and 36 weeks (panel C) and 37 weeks in those with a true GA ≥ 37 weeks (panel D). The dashed line shows perfect calibration (where observed proportion equals expected proportion), while the blue line gives the smoothed calibration slope across all pregnancies in that GA group. Average predicted risk is shown in 10 groups by predicted risk (green) and smoothed over all individuals (blue).

line of perfect calibration (see *Figure 13*, panels A and B). There was only very slight underprediction of FGR risk in pregnancies of gestational ages below 27 weeks, where the model was used to predict FGR risk at this time. Overprediction of FGR risk was evident in those with gestational ages >32 weeks, as seen in panels C and D of *Figure 13*, where the observed prevalence of FGR was much lower than predicted when using the model to predict FGR risk at 32 weeks (for those who truly gave birth between 32 and 36 weeks), or to predict FGR risk at 37 weeks (for those who gave birth at 37 weeks or later).

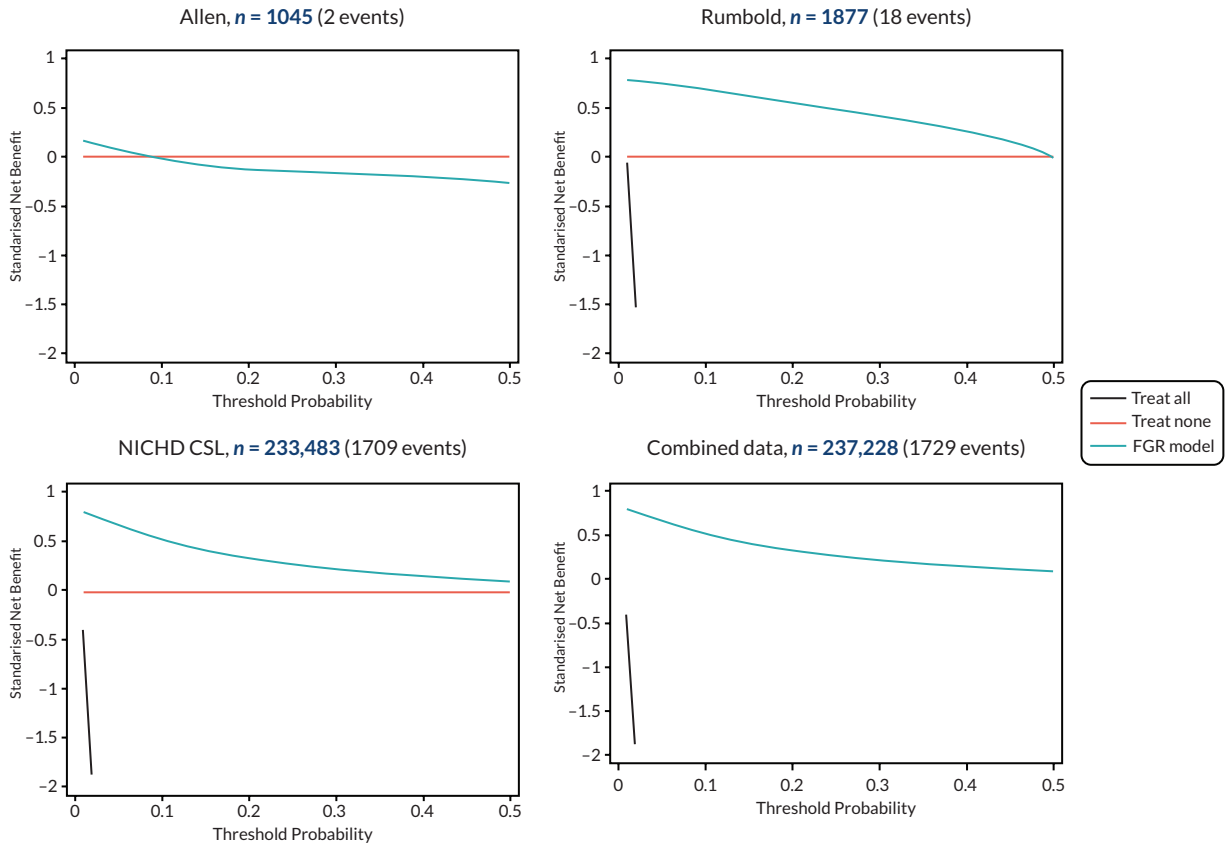

FIGURE 14 Net benefit of using the binary outcome model to predict FGR (blue) in each cohort and in the combined model development data, in comparison to treat-all (green) and treat-none (orange) strategies, as evaluated in all women (regardless of their gestational age at delivery). Decision curves are shown for threshold probabilities between 0 and 0.5, with values greater than zero implying a NB from using the model to inform decisions and those less than zero implying a net harm.

Decision curve analysis

Net benefit

A comparison of using the model to inform treatment versus treat-all and treat-none strategies was done using DCA in all participants. Calculations were conducted separately in each of the cohorts with events used for model development (Allen, Rumbold and NICHD CSL) and in the combined data from all four cohorts. NB values were multiplied at each threshold by 1000, to give the extra number of women that would be correctly treated per 1000 women for whom the model is used, with none treated incorrectly, and are presented in *Figure 14*.

A positive NB (with the curve lying above the zero line) was indicated when the model was used in the largest cohort, NICHD CSL, which was echoed in the analysis in the combined data, suggesting that a positive number of women per 1000 would benefit from being correctly identified as high risk based on the model's use than would be harmed by incorrect identification. This is true in the range of threshold probabilities from 0 up to 0.5, with the decision curve lying entirely above zero over this range. There was therefore NB from using the FGR prediction model with 'high risk' defined by cut-offs anywhere in the predefined range of probabilities from 0.01 to 0.2 considered of key interest for clinical decision-making.

The FGR prediction model also showed a positive NB in this same clinically important range in the Rumbold cohort, with a higher expected NB than was seen in the NICHD CSL cohort in the 0 to 0.5 range. The range of threshold probabilities for defining 'high risk' with a positive NB was considerably narrower for Allen, possibly a reflection of how few events there were in this cohort, with the model

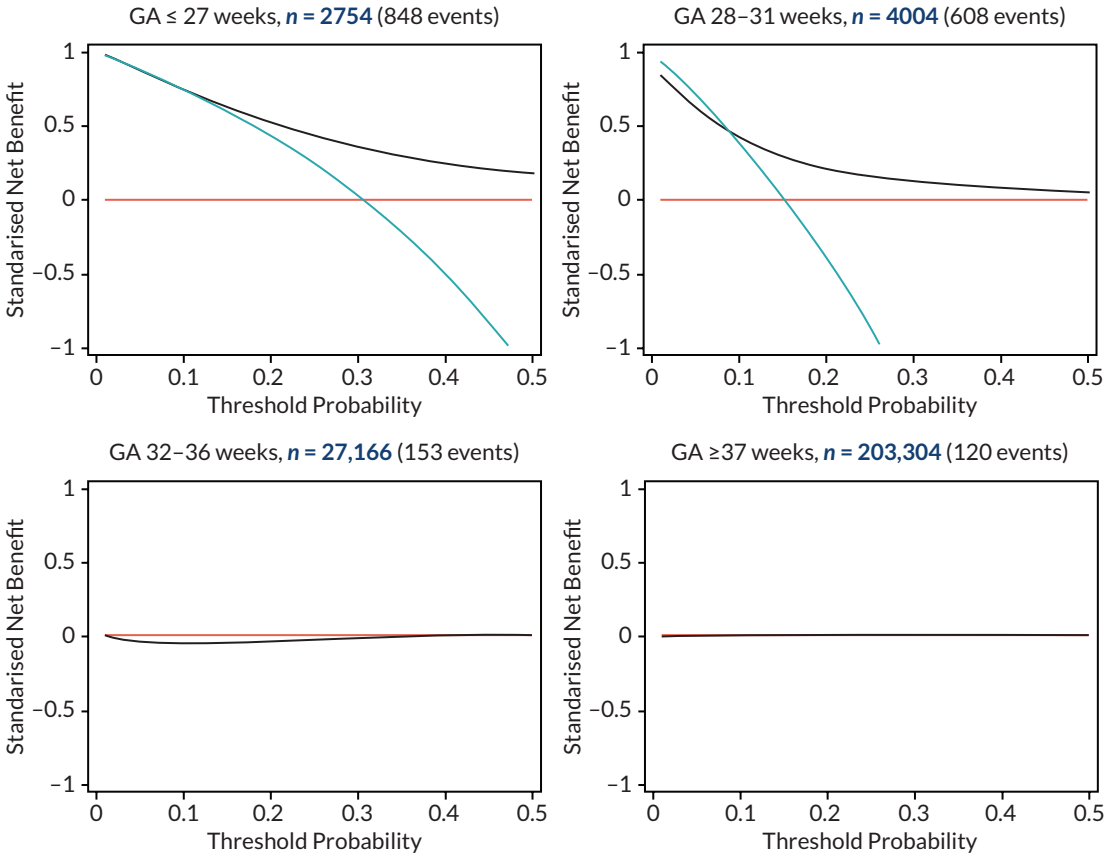

FIGURE 15 Net benefit of using the binary outcome model to predict FGR (blue) in the combined model development data (with predictions conditional on observed gestational age at delivery), in comparison to treat-all (green) and treat-none (orange) strategies, evaluated in subgroups by gestational age at delivery. Decision curves are shown for decision threshold probabilities between 0 and 0.5, with values greater than zero implying a NB from using the model to inform decisions and those less than zero implying a net harm.

becoming less favourable than a treat-none approach for threshold probabilities above 0.1, suggesting net harm when using the model in the Allen population for threshold probabilities above this range.

When considering NB separately by gestational age at delivery, decision curves show a NB in those with gestational ages before 28 weeks across the full range of threshold probabilities. A NB is further seen in those with a gestational age between 28 and 32 weeks (over and above the treat-all strategy) for all threshold probabilities >0.09 (see *Figure 15*). While no overall benefit is seen in those with gestational ages above 32 weeks, due to the extremely low prevalence of FGR in this group, NB analysis shows that using the FGR model results in no harm in these patients, while allowing substantial benefit in those who go on to deliver early. Given predictions must be generated conditional on some assumed gestational age at delivery (and true gestational age will be unknown until the time of delivery) it is important to confirm that no harm is expected in pregnancies where delivery is at later weeks.

Accuracy at specified probability thresholds

The expected number of TP, FP, TN and FN per 1000 women using the binary outcome model at different thresholds for predicted probability of FGR with complications are presented in *Table 10*. Threshold probabilities presented increase by 0.01 up to 0.2, in the range where predicted probabilities are likely to be of more interest clinically (as defined a priori by the IPPIC collaborative group), and then by increases of 0.1 afterwards.

For example, if the model was used with a threshold probability of 0.1 (10%), we would expect to identify 17 in every 1000 pregnancies as being at high risk of FGR, five of which would be expected to be truly

TABLE 10 Expected net benefit and number of TP, FP, TN and FN per 1000 women using the model at different predicted probability thresholds, based on FGR model's apparent performance in full development data

Threshold probability	TP per 1000	FP per 1000	TN per 1000	FN per 1000	Sensitivity (%)	Specificity (%)	NB per 1000
0	7	993	0	0	100	0	7
0.01	6	47	946	1	88.3	95.3	6
0.02	6	30	963	1	85.7	97	6
0.03	6	23	969	1	81.9	97.7	5
0.04	6	20	973	2	78.6	98	5
0.05	5	17	975	2	75	98.3	5
0.06	5	16	977	2	72.7	98.4	4
0.07	5	14	978	2	70.1	98.5	4
0.08	5	13	979	2	67.8	98.7	4
0.09	5	12	980	2	66.5	98.8	4
0.10	5	12	981	3	65.2	98.8	3
0.11	5	11	982	3	63.9	98.9	3
0.12	5	10	982	3	62.6	99	3
0.13	4	10	983	3	61.6	99	3
0.14	4	9	983	3	60.6	99.1	3
0.15	4	9	984	3	59.5	99.1	3
0.16	4	9	984	3	58.7	99.1	3
0.17	4	8	985	3	58	99.2	3
0.18	4	8	985	3	57.1	99.2	2
0.19	4	7	985	3	56.4	99.2	2
0.20	4	7	986	3	55.4	99.3	2
0.30	3	4	988	4	46.4	99.6	2
0.40	3	2	991	5	34.8	99.8	1
0.50	1	1	992	6	19.1	99.9	1
0.60	1	0	992	7	9.5	100	0
0.70	0	0	993	7	5.8	100	0
0.80	0	0	993	7	3.2	100	0
0.90	0	0	993	7	0.8	100	0

FGR. Of the 983 rated as low risk of FGR, we would expect to miss three who would truly be FGR babies. The corresponding sensitivity of the model used with this threshold to identify those at risk of FGR would be 65.2%, with a specificity of 98.8%. This corresponds to a NB of three in every 1000 pregnancies where model risk predictions over 10% were used to determine a pregnancy as being high risk.

There is a positive NB expected per 1000 pregnancies for all threshold probabilities below 0.9. When rounding to whole numbers, we see a NB of at least one pregnancy per 1000 for threshold probabilities below 0.53 (with predicted FGR risk below 53% from the model).

TABLE 11 Model coefficients for the final IPPIC-birthweight model, and coefficients from the models from each IECV cycle, with study-specific intercepts

	Continuous outcome model, coefficient			
	Full data	**Excluding Allen**	**Excluding Rumbold**	**Excluding STORKG**
Gestational age at delivery (weeks)				
wks^{-2}	24,200,000	24,300,000	24,600,000	24,200,000
wks^{-2} ∗ ln *wks*	−9,274,661	−9,278,365	−9,383,827	−9,256,070
Mother's weight, kg	2.708811	2.706337	2.702486	2.712717
Mother's height, cm				
cm^{3}	0.0000752	0.0000750	0.0000747	0.0000749
Mother's age, years	3.138301	3.142642	3.160154	3.160331
Nulliparous	−92.03335	−91.72014	−91.86725	−91.48258
Smoker	−118.0368	−118.025	−119.9714	−118.4065
Ethnicity				
White	ref	ref	ref	ref
Black	−174.2521	−174.4343	−174.5623	−174.2543
Asian	−73.23465	−71.06526	−73.76615	−72.71608
Hispanic	−9.562515	−9.582163	−10.10826	−9.50801
Mixed	−64.73106	−62.95441	−65.91944	−64.19719
Other	−60.38814	−59.96023	−60.27113	−62.70106
History of hypertension	−36.27748	−36.01454	−36.19573	−36.47012
History of diabetes	149.9896	150.6165	150.6321	150.0034
Assisted conception	−78.7545	−81.1364	−79.92499	−93.85787
Any previous PE	−84.16068	−84.04261	−83.93942	−84.10901
Any previous stillbirth	−13.59535	−13.11023	−13.67015	−13.1444
Any previous SGA baby	−481.7414	−486.7372	−481.4254	−485.4757
Intercept				
Average	9210.304	9212.973	9259.176	9200.492
NICHR-CSL	9210.304	9212.973	9259.176	9200.492
Allen	9243.410	–	9247.552	9214.733
Rumbold	9223.706	9168.198	–	9180.025
STORKG	9314.989	9262.523	9319.024	–
Heuristic shrinkage factor	0.9997	0.9997	0.9998	0.9997

Predicting birthweight

IPPIC-birthweight model

All candidate predictors included in the variable selection process were retained in the prediction model for birthweight, and regression coefficients for each are included in *Table 11*. Study-specific intercept

values were reasonably consistent across the four cohorts, implying the average birthweight was similar in each of the populations.

Conditional on other variables, increased mother's height, weight and age increased the predicted birthweight at a given gestational age of delivery, as did a history of diabetes. The presence of all the other predictors in the model reduced the expected birthweight. The biggest reduction of predicted birthweight from any single predictor was seen for 'previous SGA baby', with predicted birthweight reducing by 482 g for mothers who had a SGA baby in a previous pregnancy, even after adjustment for gestational age at delivery.

Mother's weight was selected as having a linear relationship with birthweight with an increase in predicted birthweight of 3 g/kg of mother's weight. This linear relationship is not consistent with the modelling of weight included in the Poon 2011 model (where linear, squared and cubed weight terms were included) but is unsurprising given the near-linear shape of the best-fit line in this data, as seen in *Figure 9*.

An estimate of heuristic shrinkage was calculated in each imputation for each cycle of the IECV, giving an estimate of 0.9997 across imputed datasets. This implies very little overfitting in the development data in any IECV cycle, and thus application of shrinkage to the model coefficients was not required.

Apparent model performance

Predictive performance measures were calculated separately by cohorts, with predictions calculated using study-specific intercepts. For a fair comparison to the observed birthweights, predictions were generated for the true gestational age at delivery. Study-specific predictive performance measures are presented in *Table 12*, along with pooled performance measures across cohorts.

Apparent calibration performance of the birthweight model (when applied using the average intercept) was good on average as expected, although it varied across datasets. Calibration slopes ranged from 0.884 (95% CI 0.809 to 0.960) in the Allen dataset, up to 1.043 (95% CI 0.994 to 1.092) in Rumbold, showing some heterogeneity across datasets. In two of the four datasets, the range of predictions was slightly too wide compared to observed values (too extreme for both low and high birthweights) as evident by an estimated calibration slope below 1. Overall, there was little miscalibration on average across all four datasets, with the pooled calibration slope of 0.989 (95% CI 0.881 to 1.098).

Calibration-in-the-large was close to zero in all datasets, with an average underprediction of birthweights between 13.4 g (Rumbold) and 104.7 g (STORKG). On average across datasets, CITL was only 44.4 g (95% CI −18.4 g to 107.3 g). This is reiterated by a pooled ratio of mean observed to mean expected birthweight of 1.017 (95% CI 0.967 to 1.066). Across datasets, the proportion of variation in birthweight explained by the birthweight model (R^2) ranged from 32.3% to 56.3%, which is moderate to large (see *Table 12*).

Model performance on internal-external cross-validation

To give a better representation of how the model might perform in new data, predictive performance measures were calculated on IECV. Given its large number of patients relative to the other cohorts, NICHD CSL[164] was forced to remain throughout all cycles of the IECV approach. Therefore, we did not include a cycle where a model was built without NICHD CSL. Thus, although there were four cohorts available, there were only three cycles of the IECV approach reported in the validation below.

Model coefficients for each cycle of the IECV process are reported in *Table 13*. These coefficient estimates were reasonably consistent when estimated in subgroups of just three out of the four available cohorts, likely due to being highly influenced by the NICHD CSL cohort, which contributed the majority of the observations to every cycle of the IECV.

TABLE 12 Apparent model performance by dataset for the birthweight model with average intercept, summarised across imputations

	Pooled estimate	Allen (*n* = 1045)	Rumbold (*n* = 1877)	STORKG (*n* = 823)	NICHD CSL (*n* = 233,483)	Full data (*n* = 237,228)
Calibration slope						
Point estimate	0.989	0.884	1.043	1.029	0.991	0.991
CI	0.881 to 1.098	0.809 to 0.960	0.994 to 1.092	0.952 to 1.105	0.987 to 0.994	0.987 to 0.994
Prediction interval	0.70 to 1.28	–	–	–	–	–
τ^2 (95% CI)	0.003 (0.000 to 0.040)	–	–	–	–	–
CITL						
Point estimate	44.445	33.106	13.402	104.685	31.425	31.534
CI	−18.444 to 107.333	7.070 to 59.141	−6.451 to 33.255	75.616 to 133.76	29.685 to 33.166	29.807 to 33.261
Prediction interval	−136.62 to 225.51	–	–	–	–	–
τ^2 (95% CI)	1400 (257 to 13,000)	–	–	–	–	–
Observed/expected						
Point estimate	1.017	1.012	1.005	1.036	1.011	1.011
CI	0.967 to 1.066	0.962 to 1.062	0.944 to 1.067	0.978 to 1.094	0.915 to 1.106	0.916 to 1.106
Prediction interval	0.95 to 1.08	–	–	–	–	–
τ^2 (95% CI)	0.000 (0.000 to 0.004)	–	–	–	–	–
R^2[a]						
Median (%)	46.9	32.7	47.8	45.7	56.1	56.0
Range	32.3–56.3	32.3–32.9	47.5–48.1	45.1–46.3	56.0–56.3	55.9–56.1
IQR	39.0–52.1	32.6–32.8	47.8–47.9	45.6–45.9	56.1–56.2	56.0–56.1

a Reported as median, range and IQR across imputations as R^2 cannot be summarised across imputations using Rubin's rules.

TABLE 13 Predictive performance of the developed birthweight model with average intercept in each IECV cycle: the external validation performance in each dataset, for the cycle in which it was excluded from model development

	Pooled estimate	Allen	Rumbold	STORKG
N for model development	–	236,183	235,351	236,405
N for external validation	–	1045	1877	823
Calibration slope				
Point estimate	1.002	0.895	1.065	1.038
CI	0.776 to 1.227	0.819 to 0.972	1.015 to 1.115	0.960 to 1.115
Prediction interval	−0.25 to 2.26	–	–	–
τ^2 (95% CI)	0.007 (0.001 to 0.144)	–	–	–
CITL				
Point estimate	9.720	−22.324	−33.419	86.406
CI	−154.317 to 173.756	−48.356 to 3.707	−53.363 to −13.474	57.308 to 115.504
Prediction interval	−943.23 to 962.67	–	–	–
τ^2 (95% CI)	4200 (801 to 76,000)	–	–	–
Observed/expected				
Point estimate	1.004	0.995	0.991	1.030
CI	0.938 to 1.070	0.949 to 1.041	0.935 to 1.047	0.974 to 1.086
Prediction interval	0.81 to 1.20	–	–	–
τ^2 (95% CI)	0.000 (0.000 to 0.008)	–	–	–
R^2[a]				
Median (%)	45.7	32.6	47.4	45.7
Range	32.2–47.8	32.2–32.8	47.1–47.8	45.0–46.2
IQR	32.7–47.4	32.5–32.7	47.4–47.5	45.5–45.8

a Reported as median, range and IQR across imputations as R^2 cannot be summarised across imputations using Rubin's rules.

Calibration measures are reported separately by cohorts in *Table 13*, giving the performance of the model developed in all but one cohort when validated 'externally' in that reserved cohort. Pooled performance estimates give the average performance across IECV cycles. Predictions were generated for the true gestational age at delivery, using the study-specific intercept for the NICHD CSL study population, because the majority of the development data in each cycle came from this cohort and so average intercepts across cohorts would be heavily influenced by the mean birthweight from the NICHD CSL data.

The calibration slope estimates across IECV cycles suggest some slight overfitting to the development data in each cycle; specifically, the largest study dominates the model estimation, and as a result, the remaining studies were slightly miscalibrated. The largest miscalibration was seen in the model developed in the cohorts excluding Allen, with the calibration slope upon 'external' validation of 0.895 (95% CI 0.819 to 0.972) – however, this is still only very slight when shown visually (see *Figure 16*). The pooled calibration slope across IECV cycles was 1.002 (95% CI 0.776 to 1.227), with negligible miscalibration on average when models were applied in the cohorts held out from model development.

The model developed in the cycle excluding the STORKG cohort was the best calibrated in terms of the calibration slope, at 1.038 (95% CI 0.960 to 1.115), although this model performed worst by CITL, a

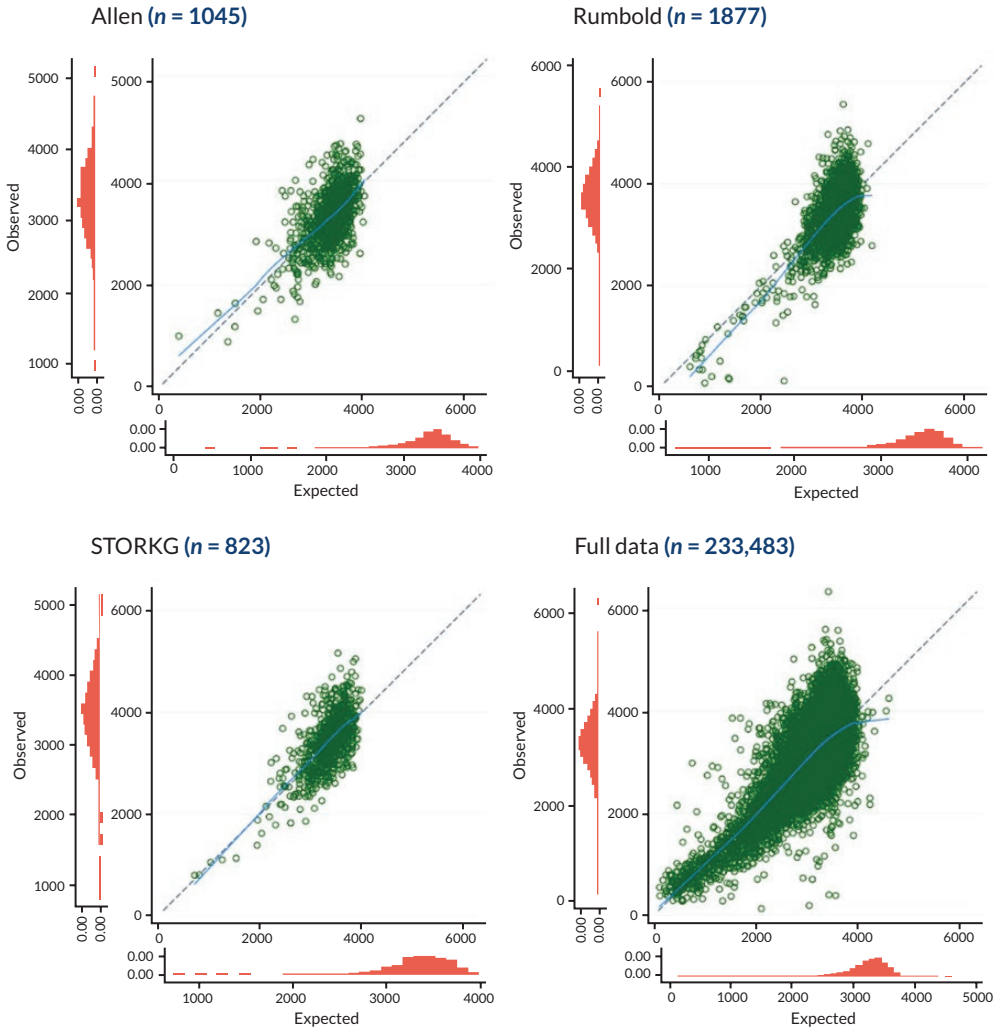

FIGURE 16 Calibration plots for the birthweight model in each IECV cycle, on external validation in the dataset excluded from the model development stage of that cycle, and apparent calibration of the birthweight model with average intercept when applied in the full dataset. Observed birthweight is plotted against predicted birthweight, with the dashed line showing perfect calibration (where observed value equals expected value), while the blue line gives the smoothed calibration slope across all pregnancies.

systematic underestimation of birthweight by 86.4 g on average (95% CI 57.3 to 115.5) in this cycle. The pooled CITL suggests an underestimation of birthweight by 9.7 g on average across IECV cycles (95% CI −154.3 to 173.8).

Predicted birthweights from the models developed in each IECV cycle (generated using the true gestational age at delivery and the study-specific intercept from the NICHD CSL study) were compared to the observed birthweights in the excluded study from that cycle. On visual inspection of the calibration plots, the smoothed calibration curves (shown in light blue) lie close to the diagonal line of perfect calibration for all cycles of IECV. Hence calibration is generally excellent, and the miscalibration noted above (in terms of calibration slope) appears to be minor.

While model calibration was good on average, miscalibration can be seen for individual observations at the higher end of the range of predicted birthweights, resulting in a wide spread of observed birthweights for a particular predicted birthweight in all cycles of the IECV. This spread is far narrower in the clinically important range of lower predicted birthweights, where pregnancies would be at higher risk of FGR.

Comparison of model performance to existing models

The Poon 2011 model performed well on average on external validation, although showed some heterogeneity in calibration across datasets, and slight miscalibration in the large. By including additional variables, on top of those from the Poon 2011 model, we hoped to reduce the heterogeneity in calibration performance across different populations. Our newly developed model is therefore referred to here as the 'updated model', as we updated the Poon 2011 model to include additional predictors.

Three of the model development cohorts were also used to externally validate the Poon 2011 model, and so predictive performance measures for the Poon 2011 and updated models were compared in these cohorts. We used the IECV performance for the newly developed model, that is, when that cohort was reserved for external validation for this comparison (see *Table 13*). Note that the final model built on all data was not represented in *Table 13*, as including the apparent performance of the newly developed model with the external validation performance of the Poon 2011 model would be an unfair comparison.

Calibration plots showing the predicted birthweight by each model (Poon 2011 and the updated model) compared to the observed birthweights in the Allen, Rumbold and STORKG datasets are given in *Figure 17*. On visual comparison, the calibration performances of both models are similar. Observed birthweights are similarly spread out for each predicted birthweight at the higher end of the range, with a narrower spread of observed values for the more clinically relevant predictions of lower birthweights. Both models appear to perform well on average, with the smoothed prediction curve (blue line) for each lying very close to the diagonal for all three cohorts.

Visual consistency in calibration plots is supported by the predictive performance statistics presented in *Table 14*, where the calibration slope, CITL and R^2 values of the two models are very similar for each of the cohorts. In particular, the R^2 values suggest that a similar amount of the variation in the observed birthweight for these three cohorts is explained by each model, despite the inclusion of new predictors (previous PE, previous stillbirth and previous FGR baby) in our updated model.

In the Allen, Rumbold and STORKG cohorts, the Poon 2011 model had a calibration slope slightly closer to one than in the updated model, but conversely CITL was improved by the updated model. Systematic underestimation of weight was seen with the Poon 2011 model for both the Allen and Rumbold cohorts, while the updated model overestimated birthweight by 22.3 g and 33.4 g on average (compared to underestimation by 64.2 g and 125.1 g) for the Allen and Rumbold cohorts, respectively. In STORKG, the updated model underestimated birthweight by 86.4 g compared to underestimation by 118.9 g when using the Poon 2011 model. The calibration slopes of the updated model and the Poon 2011 model were 1.038 (95% CI 0.960 to 1.115) and 1.009 (95% CI 0.936 to 1.083), respectively.

Model equations and summary performance measures of the developed prediction model are shown in *Table 15*. The Poon 2011 model has some miscalibration-in-the-large, the magnitude (underestimation by 64.2 g to 125.1 g, dataset dependant) of which will be more pronounced in newborns born at earlier gestations. In the updated model, the miscalibration-in-the-large (which is closer to zero than that of the Poon 2011 model) was negligible. On the whole, both models perform similarly.

Summary and example predictions

In this chapter we used IPD from the IPPIC data repository to develop two models: the first to predict the probability of FGR, defined by a birthweight below the tenth centile by gestational age with serious complications (preterm birth <32 weeks, stillbirth or neonatal death); and the second to predict birthweight at various gestational ages. Both models extend the Poon 2011 prediction model,[59] by incorporating predictors from the Poon 2011 model as a base, along with additional important predictors identified through a Delphi survey of the IPPIC Collaborative Network.

TABLE 14 External validation performance of the updated birthweight model in each IECV cycle (performance in each dataset, for the cycle in which it was excluded for model development), and the Poon 2011 model in Allen, Rumbold and STORKG

	Allen (*n* = 1045)		Rumbold (*n* = 1877)		STORKG (*n* = 823)	
	Updated model	**Poon 2011**	**Updated model**	**Poon 2011**	**Updated model**	**Poon 2011**
Calibration slope						
Point estimate	0.895	0.906	1.065	0.963	1.038	1.009
95% CI	0.819 to 0.972	0.830 to 0.981	1.015 to 1.115	0.919 to 1.007	0.960 to 1.115	0.936 to 1.083
CITL						
Point estimate (g)	−22.32	64.23	−33.42	125.06	86.41	118.92
95% CI	−48.36 to 3.71	38.39 to 90.07	−53.36 to −13.47	105.45 to 144.67	57.31 to 115.50	90.00 to 147.84
R^2[a]						
Median (%)	32.6	34.7	47.4	50.1	45.7	51.2
IQR	32.5–32.7	–	47.447.5	–	45.5–45.8	–

a Reported as median across imputations as R^2 cannot be summarised across imputations using Rubin's rules.

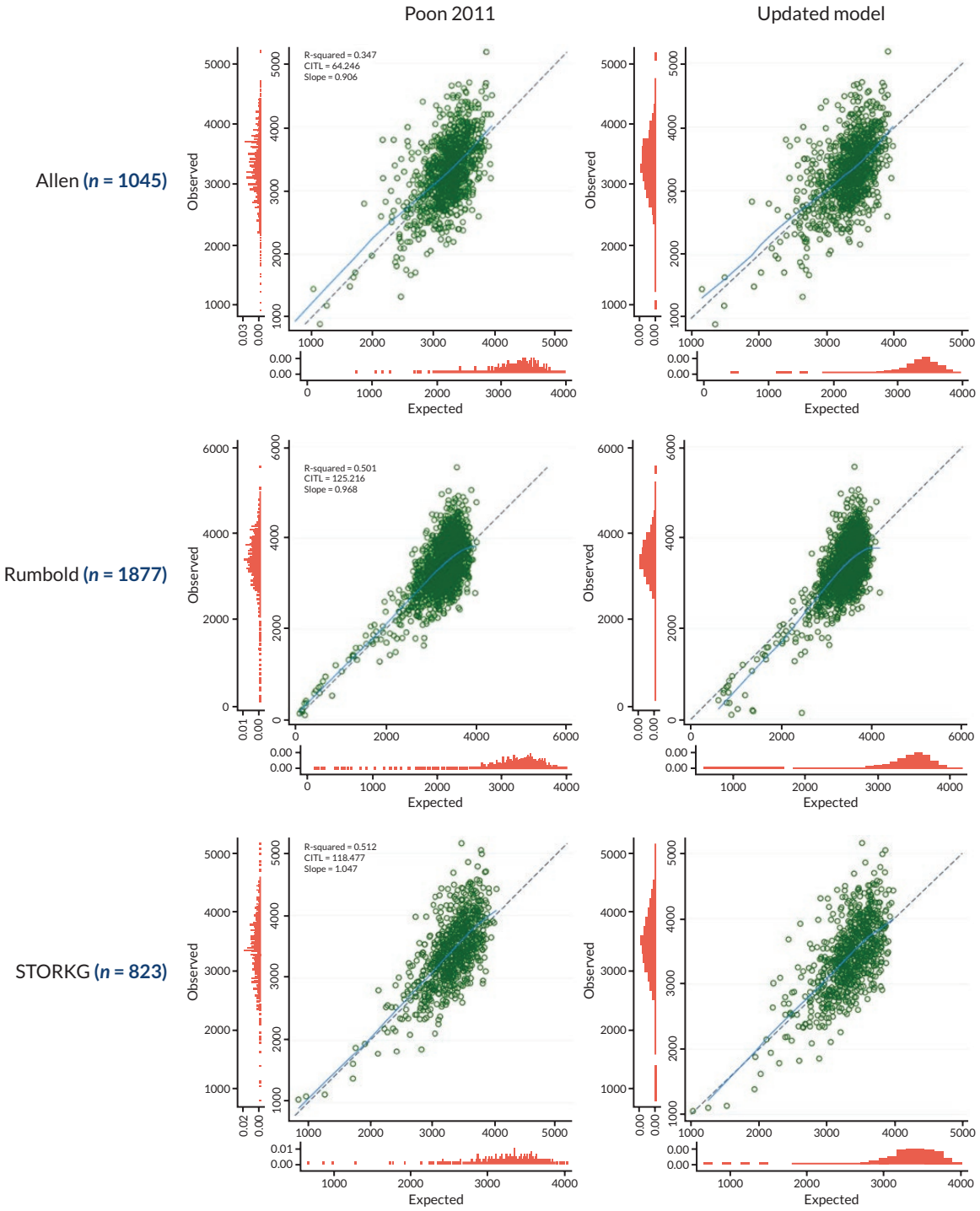

FIGURE 17 Calibration plots for the updated birthweight model in each IECV cycle (performance in each dataset, for the cycle in which it was excluded for model development), and the Poon 2011 model in Allen, Rumbold and STORKG. Plots are from a single representative imputation. Observed birthweight is plotted against predicted birthweight, with the dashed line showing perfect calibration (where observed value equals expected value), while the blue line gives the smoothed calibration slope across all pregnancies.

Both models can be used to generate predictions conditional on some assumed (clinically relevant) gestational age for delivery (or ideally a range of assumed values), as the true delivery time would be unknown at the moment of prediction. When used in combination, these models can give unique estimates of predicted birthweight and risk of FGR across the whole range of possible gestational ages at delivery. This is illustrated for two hypothetical babies in *Figure 18*, one which is clearly high risk and one that is low risk. Such plots of predictions allowing clinicians and patients to assess risks over

TABLE 15 Model equations (with average intercept) and performance summary

Outcome	Model equation	Average statistic (95% CI) [95% prediction interval]		
		c-statistic	Calibration slope	CITL
FGR (SGA with serious complications)	Logit(p) = −22.811 + 0.01 × (age) + 0.358 × (nulliparous) + 0.293 × (smoked) + 0.432 × (black) + 0.181 × (Asian) + 0.296 × (Hispanic) + 0.953 × (mixed) + 0.003 × (other) + 0.313 × (hypertension) + 0.887 × (previous PE) + 0.436 × (previous stillbirth) + 2.166 × (previous SGA baby) − 0.000000108 × (height^3) + −56,010.23 × (GA^{-2}) + 21652.92 × [GA^{-2} × ln(GA)]	0.962 (0.508 to 0.998)	0.947 (0.665 to 1.230)	0.323 (−1.881 to 2.527)
Birthweight	Birthweight = 9210.3 + 3.1 × (age) − 92 × (nulliparous) − 118 × (smoked) − 174.3 × (black) − 73.2 × (Asian) − 9.6 × (Hispanic) − 64.7 × (mixed) − 60.4 × (other) − 36.3 × (hypertension) + 150 × (diabetes) − 78.8 × (assisted conception) − 84.2 × (previous PE) − 13.6 × (previous stillbirth) − 481.7 × (previous SGA baby) + 2.7 × (weight) + 0.0000752 × (height^3) + 24200000 × (GA^{-2}) − 9274661 × [GA^{-2} × ln(GA)]	–	1.002 (0.776 to 1.227)	9.72 g (−154.3 to 173.8)
Birthweight (Poon 2011 model[59])	$\log_{10}$ (birthweight) = −0.935219 + 0.186853(gestational age) − 0.002078 × gestational age^2 + 0.003726 × weight −0.000030 × weight + 8.820640e^{-08} × weight^3 + 0.000965 × height + 0.001466 × age) − 0.000026 × age^2 + 0.016986 × parous − 0.024867 × age − 0.021769 × African − 0.017824 × South Asian − 0.005543 × East Asian − 0.009063 × mixed − 0.020995 × hypertension + 0.03143 × diabetes − 0.004015 × assisted conception	–	0.974 (0.938 to 1.011) [0.868 to 1.081]	90.39 g (37.9 to 142.9) [−78.4 to 259.2]

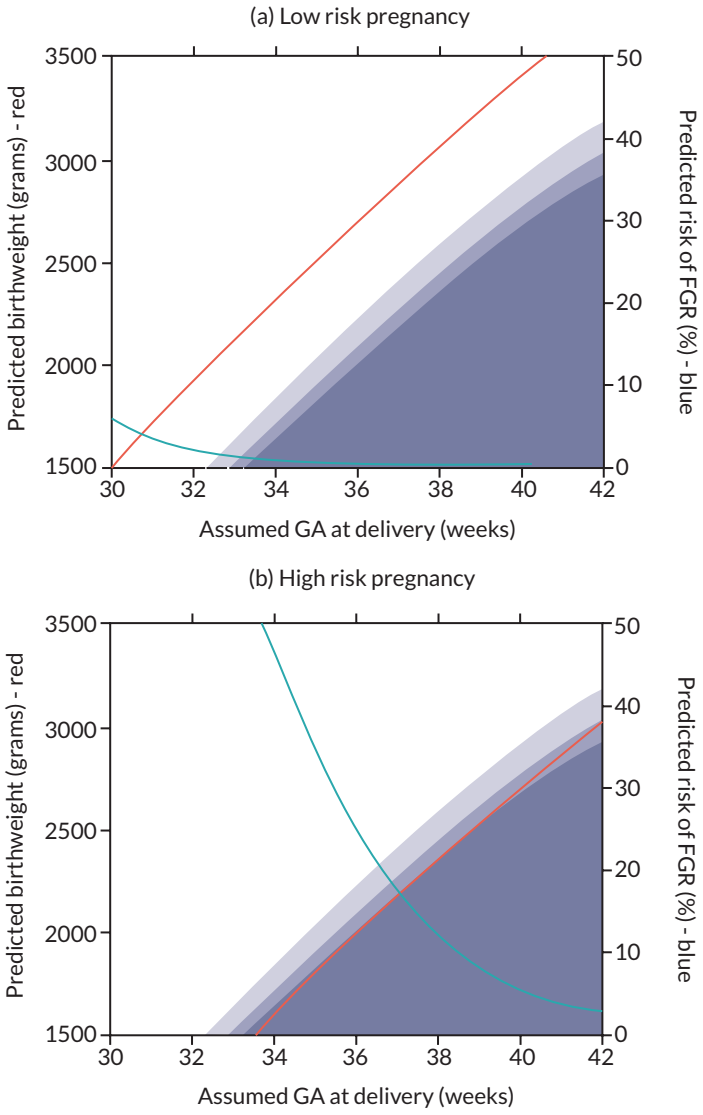

FIGURE 18 Predicted birthweight (red) and predicted FGR risk (blue) at different assumed gestational ages at delivery, using our models for two hypothetical babies (one high risk and one low risk). Shaded regions indicate birthweights below the 10th (lightest), 5th (middle) and 3rd (darkest) percentiles.

the entire range of gestational ages at delivery, and give a complete picture to enable shared decision-making around the frequency of monitoring during a pregnancy.

When these models were validated using study-specific intercepts across included IPD cohorts, the predictive performance of both models was best in the cohort contributing most to the development data (NICHD CSL).[164] For both models (with continuous and binary outcomes) the coefficients were informed mainly by this one study. Performance from IECV (for the birthweight model), and apparent performance in individual cohorts (for the FGR model) for the other three cohorts was promising, as was the performance of the Poon 2011 model for predicting birthweight when externally validated in these same cohorts.

Prediction of birthweight

Pooled performance across development cohorts was good for both continuous outcome models (the Poon 2011 model and our updated model). While the calibration slope of the updated birthweight model did not show improvement over the Poon 2011 prediction model, CITL was much improved. Miscalibration-in-the-large of the Poon 2011 ranged from 64.2 g to 125.1 g, which will be more pronounced in newborns born at earlier gestations. Therefore, there is little to choose between the

Poon 2011 model[59] and our updated model when predicting birthweight on a continuous scale. Both perform well and explain a reasonable amount of variability, especially in the lower birthweight range.

Prediction of FGR with complications

Decision curve analysis showed potential for NB of the FGR prediction model across a wide range of threshold probabilities especially in the larger included datasets, although this range was narrower in the smaller datasets. By incorporating the risk of serious perinatal complications along with birthweight centile by gestational age at delivery, the FGR prediction model offers more scope for identification of pregnancies that are at risk of adverse outcomes (over-and-above prediction of birthweight alone) which would benefit from increased monitoring. This FGR model complements the above birthweight prediction models (Poon 2011 and our updated model) and could be used in combination with these to give a full picture of the predicted extent of smallness along with the risk of FGR, conditional on different assumed gestational ages of delivery.

In summary, there is the potential for these prediction models to be useful in combination in predicting the risk of FGR and birthweight in selected populations. However, the models may need to be tailored to improve the predictive performance across settings and populations different to those included here, for example those with a different baseline risk of perinatal complications. In particular, our new FGR model would benefit from external validation to assess predictive performance in new populations, especially as our opportunity to assess 'external' performance in IECV was limited by datasets with only a few outcome events. Also, a complementary model would be useful to predict the overall risk of FGR (averaged across all potential gestational ages), to go alongside our predictions which are conditional on particular gestational ages at delivery.

Chapter 7 Costs and outcomes of IPPIC-FGR model

Objective

The main objective of this model-based health economics analysis was to compare the costs and outcomes of the IPPIC prediction models for FGR, with existing strategies in the NICE 2008 Antenatal Care guideline for monitoring FGR.[169] We employed the same strategy used by NICE by using a decision analytical model framework, due to the lack of suitable evidence and data available for use in other economic model frameworks. The previous chapter developed two prediction models (one for FGR and another to predict birthweight), with the IPPIC prediction model for FGR, being suitable for evaluation using decision analytical model framework. Any strategy for predicting FGR needs to be balanced against the resources required to deliver this strategy, within the context of finite resources of the National Health Service (NHS) and to allow for redistribution of resources more efficiently across healthcare services. We set out to provide economic evidence to help decision-makers in different healthcare settings determine which strategy provides the greatest effectiveness (perinatal death avoided) at the most reduced cost in detecting FGR.

Method

For the health economics analysis, we developed a decision tree model based on NICE 2008 model, which was the most suitable model in this case, as the individuals in the model are independent of each other, and there are no recurring events. The time horizon for the economic model is less than a year: time from earliest entry into the model to delivery of fetus is <9 months; hence no discounting is required. The model was constructed using Microsoft Excel® and compared to Strategy 1 and Strategy 3 of the NICE decision tree model as shown in *Figures 19* and *20*. The perspective adopted was that of the NHS and personal social services (PSS) as recommended by NICE;[170] and private out-of-pocket costs to women or productivity losses have not been considered for the analysis.

National Institute for Health and Care Excellence economic model strategy

There are three main branches on the decision tree of the NICE 2008 economic model, which represents three different strategies for measuring and monitoring fetal growth (with SGA being used as the proxy). The study population consists of nulliparous women with singleton or multiple pregnancies, and the outcomes are caesarean section (CS) and no caesarean section (no-CS).

Strategy 1 (no measurement of fetal growth): In this strategy, fetal growth is not measured and there is no monitoring. There are two main outcomes, the pregnant woman either goes on to deliver a baby via CS or vaginally.

In Strategy 2, the measurement of FGR is by ultrasound, and all women are offered this. It is assumed all women accept this offer. There are four possible pathways: the fetus is correctly identified as growth restricted following the ultrasound scan (TP), the fetus is correctly identified as not being growth restricted following the ultrasound scan (TN), the fetus is incorrectly identified as growth restricted following the ultrasound scan when it is within the normal size range (FP), or the fetus is incorrectly identified as not being growth restricted following the ultrasound scan when it is in fact growth restricted (FN).

In Strategy 3, the measurement of FGR is conducted using symphysis-fundal height (SFH) measurement and ultrasound. First, an SFH measurement is performed for all women. Following this, two pathways

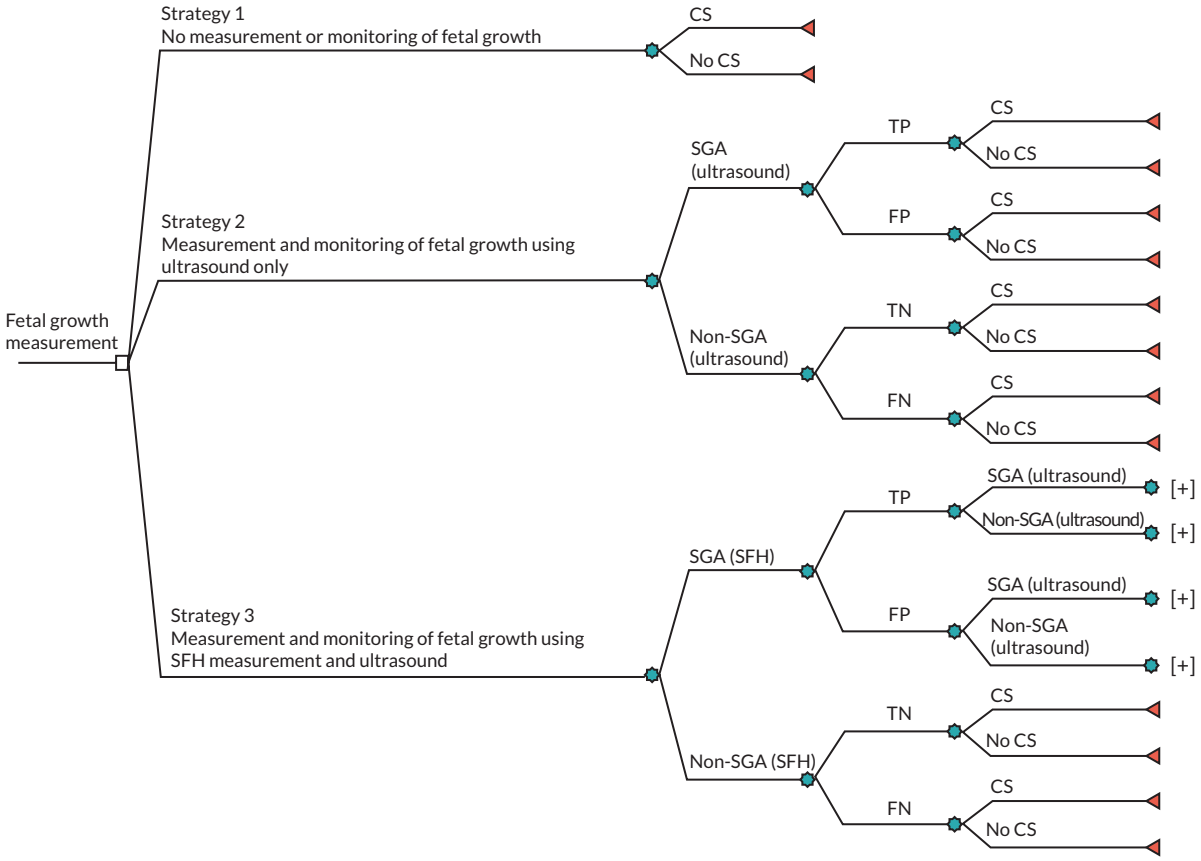

FIGURE 19 NICE decision tree for measuring and monitoring fetal growth. Strategy 1 and Strategy 3 of decision tree were compared to IPPIC prediction model strategy.

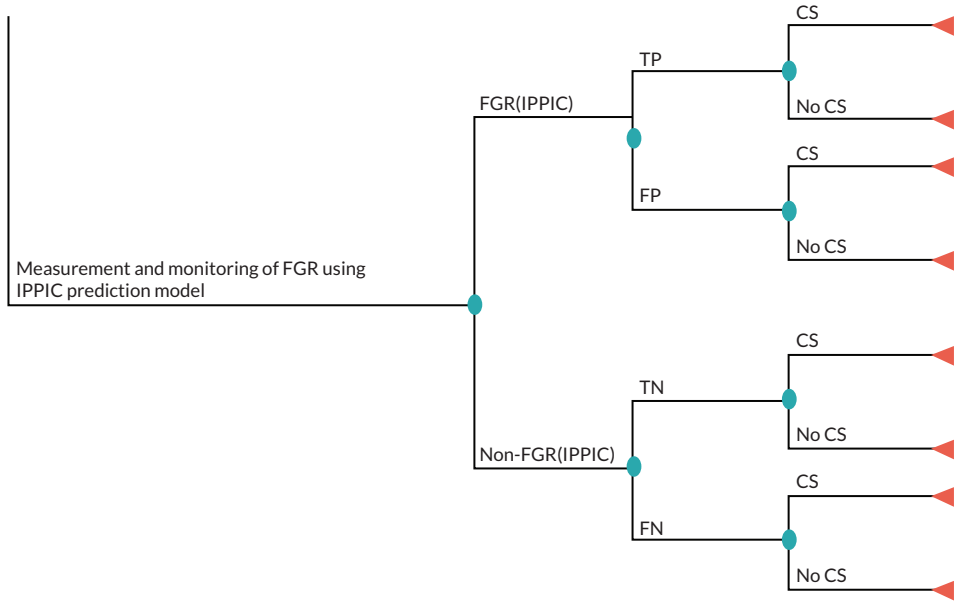

FIGURE 20 The decision tree for measuring and monitoring FGR using the IPPIC prediction model.

are possible: the baby either has FGR or no FGR. However, for each pathway, the baby could be given a correct or an incorrect diagnosis. Thus, we can have babies correctly identified as FGR (TP), incorrectly identified as not FGR (FN), incorrectly identified as FGR (FP) or correctly identified as not FGR (TN). If there is a positive identification using SFH measurement (TP and FP) an ultrasound test is performed. Once again, the ultrasound test results in four distinct possibilities (TP, FN, FP and TN).

Following each of these identifications, the pregnant woman will follow the same pathway for each branch, as in Strategy 1, that is, CS or non-CS. If women test negative for SFH (TN and FN), they are not offered an ultrasound. Their pathway is similar to Strategy 1. In the health economics analysis, we will compare a strategy using the IPPIC prediction model with Strategy 1 and Strategy 3 of the NICE 2008 economic model.

IPPIC prediction model strategy for monitoring fetal growth restriction

We compared Strategy 1 and Strategy 3 (referred to in this report as Strategy 2) of the NICE economic model with the prediction model developed to detect FGR in the earlier sections of this report. We only consider the prediction model developed for predicting FGR (Model 1), where the outcome variable is binary (FGR or no FGR). We have not considered the prediction model developed to predict birthweight where the outcome is a continuous variable (birthweight of baby) as we would have to dichotomise this variable into a binary variable (FGR or no FGR), which would be redundant, as this is captured in the binary IPPIC-FGR model. The pathway followed in this strategy starts with a baby detected with FGR or no FGR and ends with the same outcomes as the strategies in the NICE 2008 economic model (CS and no-CS).

Inputs to model

The parameters for the model were populated from multiple sources using existing literature (see *Table 16*). If parameters were unavailable, expert opinion was sought. The quality of the data source was assessed using the following criteria: data obtained from published articles, systematic reviews, meta-analyses, other economic evaluations and national registers, were considered as 'high quality'; if data were unavailable, expert opinion was sought for plausible values, or proxy data was used, and this was considered as 'low quality'. Whenever possible, all the data were from the latest available estimates. All input parameters and their quality are presented in *Tables 16–18*. The calculations for probabilities or cost values are shown in *Appendix 6*, *Table 21*.

Probabilities

Most baseline probabilities were populated using the Hospital Episode Statistics 2018–9[171] published by the NHS. Other sources of baseline probabilities include published literature as shown in *Table 16*. When

TABLE 16 Model inputs for probabilities and diagnostic test performances

Parameter	Proxy variable	Value	Source	Quality
CS		0.2951	Hospital Episode Statistics 2018–9[171]	High
Non-CS		0.7049	Hospital Episode Statistics 2018–9[171]	High
CS (FGR)		0.9000	Assumption	Low
FGR (SFH + ultrasound)	FGR	0.0300	Vieira,[174] NICE guidelines[169]	Low
FGR (prediction model)	FGR	0.0300	Vieira,[174] NICE guidelines[169]	Low
Sensitivity of SFH measurement		0.59	Pay[175]	High
Specificity of SFH measurement		0.9700	Pay[175]	High
Sensitivity of ultrasound scan		0.5556	Haragan[176]	High
Specificity of ultrasound scan		0.9598	Haragan[176]	High
Sensitivity of prediction model		0.678		Low
Specificity of prediction model		0.987		Low

Notes
Refer to *Appendix 6*, *Table 21* for indirect calculations.
High quality: Published articles, systematic reviews, meta-analyses, economic evaluations and national registers. Low quality: Data unavailable, expert opinion was sought for plausible values, or proxy data.

TABLE 17 Model inputs for costs

Parameter	Value	Source	Quality
Prediction model	£0.00	Assumption	Low
CS	£3976.91	National schedule of reference costs[172]	High
Non-CS	£2099.34	National schedule of reference costs[172]	High
SFH measurement	£4.87	PSSRU 2006[177]	High
Ultrasound fetal growth scan	£57.07	National schedule of reference costs[172]	High

Refer to *Appendix 6*, *Table 21* for indirect calculations.

TABLE 18 Model inputs for perinatal deaths averted (outcomes)

Perinatal deaths saved for every 1000 FGR babies	Value	Source	Quality
NICE Strategy 1: No measurement or monitoring of FGR	50	NICE assumption	Low
NICE Strategy 2: Measurement of FGR using SFH measurement and ultrasound	125	NICE assumption	Low
Strategy 3: Measurement of FGR using prediction model	250	IPPIC study group assumption	Low

no statistics were available, we used proxy variables in our assumptions, which were considered as low quality. Expert clinical input was obtained to verify these assumptions, so that they were as reasonable as possible, and had clinical validity. The sensitivity and specificity of the prediction model for FGR with severe complication is obtained at the threshold probability of 0.08 from section 6.3.4.2 above (see *Table 10*). This threshold probability was deemed to be the most plausible value to explain model performance. The choice for the threshold probability was guided by the desire to obtain a high enough specificity without losing out on the sensitivity.

Costs

Costs were presented in Great British pounds in 2019–20 prices. The majority of the unit costs were obtained from the National Schedule of Reference Costs: main schedule.[172] If the data was unavailable for the mentioned year, it was inflated using the NHS Cost Inflation Index pay and prices indices from the previous available years.[173] If direct cost values were unavailable, they were calculated using weighted averages, as shown in *Appendix 6*, *Table 21*. The cost of the prediction model was assumed to be £0, as there is no additional associated cost that needs to be employed by the healthcare provider for its use. However, additional cost for ultrasound scan was assigned to those predicted to have FGR using the prediction model (TP and FP), as this would reflect what would happen in clinical practice if the prediction model was introduced.

Outcomes

Assumptions have been made regarding how many perinatal deaths can be prevented for every known 1000 FGR fetuses using each strategy. About 60,000 SGA babies are born every year, of which just under a third (approximately 18,050) are expected to have FGR.[169] Half of these babies will not survive, regardless of intervention provided. Of the remaining, we can expect about 10% (903) babies to survive when using Strategy 1 (no testing), around 25% (2256) babies survive when using Strategy 2 (SFH + ultrasound) and around 50% (4513) babies survive when using Strategy 3 (prediction model). In other words, for every 1000 known FGR fetuses, approximately 50 perinatal deaths can be prevented

using Strategy 1, 125 perinatal deaths can be prevented using Strategy 2 and 250 perinatal deaths can be prevented using Strategy 3.

Analysis

Costs and outcomes analysis

For the base-case analysis, we used a decision analytical model using a deterministic approach to compare the costs and outcomes from an NHS and PSS perspective. Costs are in 2019–20 prices and outcomes are presented as the number of the perinatal deaths avoided.

Sensitivity analyses on cost for model performance

A sensitivity analysis was conducted by varying the performance of the prediction model. In our base-case scenario, we considered the model sensitivity to be 67.8% and the model specificity to be 98.7%. However, we wanted to know how the results would change, if the prediction model gave perfect information (i.e. 100% sensitivity and 100% specificity) and if it performed only as good as a coin toss (i.e. 50% sensitivity and 50% specificity).

Resource impact assessment

We also sought to conduct a resource impact assessment. This formal assessment helps to inform and quantify the costs or savings expected from implementing that guideline.[178] This cost and savings may be in the form of cash or non-cash impact for both providers and commissioners. An example of a non-cash impact is improving capacity building that does not result directly in saving money. Impact assessment also incorporates the changes in costs and savings related to changes in number of staff being employed, necessary staff training, changes in facilities required, changes in patient flows and changes in demand of the service. In our study, we wanted to assess what the resource impact would be to implement a national guideline of using a prediction model to identify FGR. However, before conducting the formal resource impact assessment, it was imperative to conduct an Evaluability Assessment.[179] This would address whether an impact assessment is required in the first place, whether it is expected to provide additional information, and if required, what would be the most appropriate methodology to do so.

Results

Base-case analysis

For the base-case analysis, we have ordered the different pathways in terms of increasing cost. Compared to a strategy of no testing for FGR, Strategy 3 of using a prediction model costs £1880.61 more per 1000 babies. On average, this slight increase in cost is associated with the largest number of perinatal deaths avoided (see *Table 19*). When the prediction model was compared with screening

TABLE 19 Base-case costs and outcomes results

	Expected cost per 1000 FGR babies	Perinatal deaths saved for every 1000 FGR babies	Incremental cost	Incremental effect
NICE Strategy 1: No measurement or monitoring of FGR	£3,032,433.03	50		
Measurement of FGR using prediction model	£3,034,313.64	250	£1880.61	200
NICE Strategy 2: Measurement of FGR using SFH measurement and ultrasound	£10,106,986.75	125	£7,072,673.11	−125

using only SFH and ultrasound (NICE Strategy 2), the model was cheaper and again more perinatal deaths were prevented.

Sensitivity analysis

The results of the sensitivity analysis are shown in *Table 20* and are in line with the base-case results. At all different levels of model performance (100% sensitivity and 100% specificity; 50% sensitivity and 50% specificity), Strategy 3 of using a prediction model to predict FGR prevented more deaths than the two NICE strategies.

Results of resource impact assessment

At the Evaluability Assessment stage, we found that we do not expect additional resource impact in addition to the impact associated with what has been described in the health economics analysis. This is because, we do not expect there to be any further costs associated with either hiring more staff or training staff to use the prediction model. Staff at hospitals are already trained to use risk assessment tools as part of their antenatal care practice. No additional facilities are required, nor do we expect changes in patient flows in the next 5 years. Thus, it was decided that a specific resource impact assessment was not needed for the prediction model.

Limitations of the economic analysis

Since development of the economic model by NICE in 2008, clinical care and guidelines have changed significantly. In addition to the 2013 Royal College of Obstetrics and Gynaecology (RCOG) green top guideline, there is also the Saving Babies Lives Care Bundle for assessment of FGR risk and the American College of Obstetricians and Gynecologists (ACOG) guidelines, which proposes alternate strategies for screening for FGR. In addition, the RCOG have a draft update to their 2022 guideline, which is expected to result in further changes in screening for FGR. The definition of FGR also varies, with some guidelines defining as small babies <10th centile (with or without customisation of centiles). A full economic evaluation is needed that takes these into account, which utilises trial or observational study data, considers quality-adjusted life-years (QALYs) to inform the cost-effectiveness of monitoring of FGR in both high and low-risk pregnancies. Hence, our economic evaluation using NICE 2008 model and care pathways may not be generalisable to current clinical practice.

TABLE 20 Sensitivity analysis of the costs and outcomes

	Expected cost per 1000 FGR babies	Perinatal deaths saved for every 1000 FGR babies	Incremental cost	Incremental effect
When sensitivity = specificity = 100%				
NICE Strategy 1: No measurement or monitoring of FGR	£3,032,433.03	50		
Measurement of FGR using Prediction model	£3,034,145.27	250	£1712.24	200
NICE Strategy 2: Measurement of FGR using SFH measurement and ultrasound	£10,106,986.75	125	£7,072,841.48	−125
When sensitivity = specificity = 50%				
NICE Strategy 1: No measurement or monitoring of FGR	£3,032,433.03	50		
Measurement of FGR using Prediction model	£3,060,970.40	250	£28,537.37	200
NICE Strategy 2: Measurement of FGR using SFH measurement and ultrasound	£10,106,986.75	125	£7,046,016.35	−125

We did not define the time point of entry into the model in terms of gestation in weeks. This may mean that resource use such as antenatal clinic visits, ultrasound scans or further tests and investigations may not have been included and there may be an under-representation of the true cost of the pathways. Assumptions on the mortality rate of FGR fetuses were based on estimations from the 2008 NICE economic model which may have resulted in exaggerated numbers of avoidable perinatal deaths across all strategies. The impact of this is more likely on the outcomes, rather than which strategy is the most cost-effective. The main outcome was in terms of the number of perinatal deaths avoided and was based on assumptions due to the lack of data. Decision-makers such as NICE prefer the final outcome to be in the form of a QALY, so then an incremental cost-effectiveness ratio (ICER), namely the cost per QALY gained can be estimated. This ICER can then be used to compare across different diseases and different interventions and allow decision-makers to make efficient choices when resources are scarce.

We conducted a simple deterministic model using point estimates as we did not have the necessary information such as Cis required for probabilistic modelling and to capture the true uncertainty around the model. The model utilises a short-term horizon only. Data were not available to populate the model, for example on the increased risk of complications such as neurodevelopmental delay at two years and the increase in risk of adult-onset diseases in infancy such as obesity, type 2 diabetes and cardiovascular disease.

Summary

Our economic analysis suggests that compared to strategies of no screening for FGR and measurement of FGR using SFH and ultrasound, based on the NICE 2008 model, there is potential for a strategy of using IPPIC-FGR model followed by ultrasound to prevent perinatal deaths. Sensitivity analyses conducted changing the model performance were in line with base-case results. The costs and outcomes analysis carried out using the NICE model is not reflective of the complex variation in current practice. The findings presented here will benefit from verification in well designed and conducted research studies with a full economic evaluation.

Chapter 8 Discussion

Summary of the findings

The newly developed and validated IPPIC-FGR model (FGR probability of FGR at various assumed gestational ages at delivery), and the updated and validated IPPIC-birthweight model accurately predict the probability of FGR (birthweight < 10th centile *and* preterm birth < 32 weeks, stillbirth or neonatal death) and birthweight, respectively, for various assumed gestational ages at birth. The IPPIC models have minimal miscalibration and excellent discrimination. Of the previously published models, the Poon 2011 birthweight model, which was used to update the IPPIC-birthweight model, has good calibration when validated in IPPIC cohorts. The IPPIC-FGR model shows net clinical benefit over a wide range of predicted probability thresholds. Its use is more cost-effective than alternate screening strategies for FGR.

Strengths and limitations

Our work complements the ongoing national efforts to reduce stillbirth and adverse perinatal outcomes, for which undiagnosed FGR is a major risk factor. Our IPD meta-analysis is the first to simultaneously develop and validate the performance of prediction models for FGR and birthweight. We used an unambiguous definition for FGR as our main outcome. By including both SGA, with severe complications such as stillbirths and neonatal deaths, we aimed to identify those babies at maximum risk of adverse outcomes, and not small but healthy babies. We accounted for treatment paradox of early delivery of a FGR baby preventing stillbirth or neonatal death,[48] by including preterm delivery before 32 weeks as a component of the outcome. By keeping our predictions on the continuous scale in our IPPIC-birthweight model for various gestational ages at delivery, we were not limited by arbitrary cut-offs used to define FGR or SGA. Such an approach also allows clinicians to calculate predicted centiles using any fetal growth standard of choice (e.g. GROW, INTERGROWTH 21st, WHO).[50,51]

Both IPPIC models can be used to generate predictions conditional on any assumed (clinically relevant) gestational age for delivery (or ideally a range of assumed values), as the true delivery time would be unknown at the moment of prediction. When used in combination, these models can give unique estimates of predicted birthweight and risk of FGR across the whole range of possible gestational ages at delivery, allowing patients to contribute to shared decision-making with clinicians around the frequency of monitoring during a pregnancy.

The IPPIC Network is collaborative in nature and was established with data provided by leading researchers with shared interest in the prediction and prevention of pregnancy complications.[44] By sharing their study data, these individuals have displayed buy-in to the research objective, which will help promote application of the developed prediction models in clinical practice. Use of this repository of cleaned, standardised and quality-assessed data from multiple cohorts increased the power of the study beyond what is achievable in a single primary study, minimised the potential for model overfitting and enabled the development and validation of robust prediction models.

We carried out a systematic approach to prediction of FGR, by first identifying existing prediction models, followed by external validation within individual IPPIC cohorts to assess transportability of identified prediction model to different populations and settings. Our model development work built on existing prediction models that showed promising performance, by informing candidate predictor selection. Clinical input was also used to prioritise predictors considered for model development. Multiple imputations were used to handle missing data for both predictors and outcome to avoid the loss of useful information.[60,180] We used rigorous statistical methods to develop the prediction

models and assess their accuracy, including undertaking a formal internal and external validation within the IPD cohorts. Predictors included in the final models are those that are clinically relevant and routinely available in both low and high-resource settings. We based our economic modelling on the NICE economic model for monitoring fetal growth published as part of NICE Antenatal care guideline 2008.[21]

There are some limitations to our study. Most of the published models identified to predict fetal growth or birthweight could not be externally validated due to differences in outcome definition reported by study authors, or because they included predictors measured late in pregnancy, and as such were more relevant for diagnosis than prediction of fetal growth.[26] We were also unable to validate eight prediction models that included predictors not available in any of the IPPIC cohorts. The use of data from existing studies for external validation of prediction models using IPD meta-analysis, is limited by variation across studies in whether and how relevant participant characteristics (as potential predictors) are recorded in these studies. However, IPD meta-analysis still provides the best opportunity to validate existing prediction models across multiple studies. Primary studies will require significant resources in order to accomplish what can be done using IPD meta-analysis from existing studies, especially with regards to generalisability, where multiple primary studies will be needed to validate prediction models. Some studies reported the development of various prediction models using data from the same cohort of women, with each subsequent publication assessing the addition of a new candidate predictor. This hinders the identification of published prediction models for external validation, and artificially increases the number of developed prediction models for fetal growth. Internal external validation of our model for FGR was limited by too few outcome events in some of the individual IPPIC cohorts. Our IPPIC-FGR prediction model was better calibrated in pregnancies with gestational age <32 weeks, however this is expected considering delivery at <32 weeks is part of our composite definition of FGR.

Our health economics analysis relied on data and structural assumptions for the decision tree model, which come with uncertainties. We however utilised high-quality data sources as much as possible such as from published meta-analysis and other economic models to inform our input and assessed the quality of all input parameters in a transparent way. This analysis was based on the NICE 2008 economic model and is not reflective of the complex variations in current practice. A detailed full-scale economic evaluation is needed, which evaluates the various strategies for risk assessment of FGR currently in use in management of pregnancies at risk of FGR. Ideally, with health economics models we can compare different interventions using outcomes based on robust clinical and health-related quality of life (HRQoL) data such as the QALY. However, due to the lack of reliable clinical and HRQoL data, the IPPIC IPD not being a primary study and consisting of studies none of which reported QALYs, our model does not follow the same structure. Instead, we take a similar approach to the NICE model,[169] and consider the number of preventable perinatal deaths attributable to FGR for the strategy to be more effective for the measurement and monitoring of fetal growth.

Our models also require the user to enter the assumed gestational age at delivery. While the expected date of delivery is not known when making a prediction, entering various possible gestational ages for delivery allows the user to produce a plot of birthweight predictions across various time points. This was illustrated at the end of *Chapter 6*. In further research, a complementary model would be useful to predict the overall risk of FGR (averaged across all potential gestational ages), to go alongside our predictions which are conditional on gestational ages at delivery.

Although planned, we were unable to assess the performance of the models by population and trimester of use due to heterogeneity in reporting of population characteristics and paucity of data on onset of FGR. Our final model only included clinical predictors, so we did not compare performance based on choice of predictors (ultrasound and biochemical markers) for predicting FGR or birthweight. It is possible that addition of further predictive markers could have improved the performance of the IPPIC models.

Comparison to existing evidence

Until now, no individual test is satisfactorily predictive of FGR to warrant recommendation in routine clinical use.[22] There is considerable variation between guidelines on screening for FGR. The UK RCOG guideline provides a list of arbitrarily categorised 'major' and 'minor' risk factors based on clinical history, and recommends regular ultrasound for women with one 'major' or three or more 'minor' risk factors.[11] The ACOG recommends screening for unspecified medical and obstetric risk factors, but does not recommend use of uterine artery Doppler or biochemical markers, citing lack of evidence on improvement of outcomes.[12] The Society of Obstetricians and Gynaecologists of Canada calls for clinical risk factors-based screening, without specifications on what these are,[13] while the Royal Australian and New Zealand College of Obstetricians and Gynaecologist suggests risk assessment through a combination of biomarkers, Doppler ultrasound and 'major' maternal clinical risk factors.[14] The choice of risk factors and their combination to predict FGR in any of the above guidelines is not based on formal predictive modelling. Their accuracy in predicting FGR is also not known.

Existing prediction models have predicted risk of SGA fetus as a surrogate measure for infants at risk of FGR,[181] with variously defined cut-offs for FGR, limiting the power and usefulness of the prediction model, by not linking these birthweight cut-offs to serious perinatal complications such as stillbirth, neonatal death, extreme preterm birth or birth trauma.[38] These SGA prediction models have mostly never been externally validated, and those that have been independently validated report limited predictive performance.[16,159] As such none are recommended for routine clinical use. None of the models identified in our search predicted our predefined outcome for FGR (SGA with serious complications of stillbirth, neonatal death or preterm birth before 32 weeks' gestation), and we were only able to independently validate one birthweight model[59] which showed slight overfitting in the validation cohorts. Individual calibration of the model across the different IPPIC cohorts was good, with moderate heterogeneity in calibration performance between the cohorts. Underprediction of birthweight by the model was consistent across different gestational ages of delivery, and this would have more of a relative impact on predictions for babies born at earlier gestational ages where expected birthweight is lower. Heterogeneity was also high across the different gestational age groups with wide prediction intervals.

The IPPIC-FGR and IPPIC-birthweight models extended the Poon 2011 birthweight prediction model[59] by building on predictors included in the original model. The IPPIC-birthweight model had good summary predictive performance, with only slight evidence of miscalibration in calibration performance, and minimal overprediction of birthweight. It underestimated the birthweight by less (12.9 g to 17.2 g) in the validation cohorts, compared to underestimation of birthweight of 64.2 g to 125.1 g by the Poon 2011 birthweight model. The IPPIC-FGR model had good summary predictive performance, with discrimination and calibration slope near 1. The model was also clinically useful across a wide range of predicted threshold probabilities, covering the identified range of 0.01 to 0.2 considered to be of interest for clinical decision-making.

There is scarce empirical evidence on the cost-effectiveness of screening for FGR. A recent study estimating the cost-effectiveness of universal routine screening by ultrasound for fetal growth reported that this was unlikely to be cost-effective.[182,183] Our health economics analysis of the IPPIC model to predict FGR, built upon the previously published economic model structure and care pathways for monitoring fetal growth by NICE 2008.[21] The NICE economic evaluation showed that although there was poor evidence on clinical effectiveness of monitoring of fetal growth by ultrasound or using SFH and ultrasound, these strategies were cost-effective compared to no screening.[21] Although our economic analysis of the IPPIC model showed that the model prevented more perinatal deaths than strategies of no screening for FGR or measurement of FGR in all fetuses using SFH and ultrasound scan, the screening strategies used in NICE 2008 are not reflective of current practice.

Relevance to clinical practice

Prediction of FGR allows for early identification of women at increased risk of FGR, who may benefit from closer monitoring in pregnancy or preventative interventions such as early administration of aspirin. Any effort to prevent adverse perinatal outcome will need to identify pregnancies that are at risk of delivering a growth-restricted baby with severe complications to assess the severity of smallness, determine the timing and frequency of surveillance and plan timing and mode of delivery. Current approach to screening differs by country and is mostly based on use of individual clinical risk factors to assess risk of FGR, which has been shown to have minimal predictive performance.[16] Our study combined clinical characteristic predictors in mathematical models to provide accurate FGR risk prediction, which have good performance when externally validated in different cohorts, looking across all observed gestational ages at delivery. Only clinical characteristic predictors are included in both IPPIC models, which make them applicable to both low and high-resource settings. The predictors included are easy to measure and routinely available in clinical practice. The prediction models will be particularly useful when used in combination to predict the risk of FGR and birthweight, as together they provide more scope to identify pregnancies that are at high risk of adverse outcome in addition to the birthweight at various gestational ages, which can inform decision for closer monitoring or intervention. Incorporating the IPPIC-FGR prediction models in practice will be straightforward as no additional measures are required to calculate the risk of FGR. However, we need to make sure that resources such as staff time and any training needs associated with using the prediction model have been costed appropriately. By working closely with clinical academics involved in the development of the RCOG Green Top national guideline on SGA fetus, and the RCOG fetal medicine clinical study group, we aim to facilitate their incorporation within national and international recommendations.

Relevance to research

FGR continues to be a research priority area. Our work is in direct response to the call of NICE guidelines and RCOG for predictive tests or strategies to identify women at risk of small baby, particularly for growth-restricted infant with complications,[11,21] and the priorities of the Department of Health to reduce stillbirths and neonatal deaths. By developing models that predict the risk of developing a growth-restricted baby with serious complication, as well as the extent of its smallness, our models provide comprehensive information to help plan management. Also, a complementary model would be useful to predict the overall risk of FGR (averaged across all potential gestational ages), to go alongside our predictions which are conditional on particular gestational ages at delivery. Further research is needed on the implementability of the IPPIC models in routine clinical practice and to determine any barriers and facilitators to its use. This should include assessment of the acceptability of the prediction models as screening tools for pregnant women and their families, as well as healthcare providers. Research is also needed to identify the acceptable care pathways for various predicted risk thresholds, and this should involve relevant stakeholders, such as healthcare providers, pregnant women and their families. The impact of using the IPPIC-FGR prediction models in clinical practice may require evaluation through cluster-randomised trials to assess whether use of the models improves perinatal outcomes. Such a trial could evaluate use of the models to inform interventions (close monitoring or planned delivery) compared to routine care on perinatal mortality and morbidity. Although feasibility of such a trial is questioned due to the sample size required to show effect on perinatal mortality, such a study might look at proxies of perinatal mortality such as morbidity to achieve sufficient power.[183]

Using IPD meta-analysis for the development and validation of the IPPIC prediction models has provided us with an increased sample size beyond any individual study, and more diverse populations for inclusion in our research. Using data from across the IPPIC data repository allowed us the opportunity for broader validation of models across different settings, populations and subgroups of interest, although this was limited due to the availability of predictor variables across the different cohorts. Primary studies on outcomes in pregnancy should collect information on fundamental predictors to minimise the

impact of systematically missing data in their study when considered for use in IPD-based projects.[184] Despite there being more than 4.5 million pregnancies from 94 cohorts contained within the IPPIC data repository, the absence of fundamental predictors restricted the number of cohorts that could be used for model development and limited the number of existing models that could be externally validated.

A key problem in the prognostic field is that many prediction models are being developed, with far fewer externally validated.[185,186] While we attempted validation of all existing prediction models for FGR and birthweight, our ability to do so was limited by the lack of consistent predictor variables reported across the IPPIC cohorts, as well as the inclusion of model predictors that are rarely recorded in practice. While novel methods for multiple imputation can be used to account for systematically missing data (accounting for both the clustering of participants within studies and heterogeneity between studies), such methods can substantially complicate analyses and increase required computation time. The best approach to handle such data is often context specific, and in our study, there were issues of convergence of imputation models when we tried to impute for systematically missing predictors. We therefore decided it was better to impute within each study separately, which naturally retains the clustering of participants within studies and any heterogeneity between studies, though at the expense of not allowing systematically missing predictors to be imputed. The methods and software available for systematically missing data are constantly evolving, and while improved recording of core predictors at the primary study level would vastly reduce the need for such approaches, future IPD projects should also stay abreast of advances in methodology in this area, to ensure they maximise the use of the available data.

Our IPD meta-analyses allowed us to explore predictive performance more extensively than a single validation study. This is important, as calibration and discrimination performance of prediction models are known to vary across populations, which can clearly be seen in our external validation of the Poon 2011 model and the IECV of the IPPIC-birthweight model. This leaves a challenge for those wanting a single model for use everywhere, especially given models in this context appear to underpredict birthweight in some populations while overpredicting in others. It may be that locally recalibrated models may be a better way forward, where IPD gives us the ability to update and tailor these models to improve performance in specific settings and allow the same base model to be accurately fitted to multiple populations.[187]

Researchers should adhere to recommended practice guidelines during economic evaluation of prediction models and follow approved methods for data collection, analysis and modelling. When including a prediction model in an economic model, key steps should be included, such as structure of the tree or model (to define possible pathways), estimate probabilities of the different pathways, assign values to costs and outcomes including any assumptions, analyse or roll back the tree, and explore any uncertainty in the model. Tools such as the CHEERS (Consolidated Health Economic Evaluation Reporting Standards)[188] or Philips's checklist[189] should be followed to ensure that the health economic evaluation being carried out is consistently and transparently reported, and that good practices are followed when developing economic models to better inform health decisions.

Conclusion

The IPPIC-FGR and IPPIC-birthweight models accurately predict the risk of FGR and birthweight for various assumed gestational ages of delivery. The latter has better calibration performance than existing model. IPPIC-FGR model has clinical utility across wide probability thresholds and is more effective compared to alternate strategies of screening for FGR using SFH and ultrasound. Use of the IPPIC models in combination has the potential to identify women at high risk of FGR and assess the severity of smallness of fetuses across the range of potential gestational ages at delivery to plan appropriate management and minimise adverse perinatal outcomes.

Acknowledgements

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

We acknowledge NICHD DASH for providing the NICHD CSL study data that was used for this research, the contribution of the Principal Investigator(s) who conducted the original study from which the data were generated, and NICHD and other funding organisation, if applicable, that supported the original study. We also acknowledge Professor Gordon Smith for providing the Pregnancy Outcome Prediction study data which contributed to the external validation of existing prediction model component of the project.

We are thankful to members of the Independent Steering Committee, which included Professor Gary Collins (Chairperson, University of Oxford), Dr Chua Mei Chien (KK Women's and Children's Hospital), Professor Gita Mishra (University of Queensland), Professor Fiona Denison (University of Edinburgh), Mrs Rebecca Harmston (Katies Team) and Professor Marian Knight (University of Oxford) for their guidance and support throughout the project.

The following are members of the IPPIC Collaborative Network:

Alex Kwong – University of Bristol; Ary I Savitri – University Medical Center Utrecht; Kjell Åsmund Salvesen – Norwegian University of Science and Technology; Sohinee Bhattacharya – University of Aberdeen; Cuno SPM Uiterwaal – University Medical Center Utrecht; Annetine C Staff – University of Oslo; Louise Bjoerkholt Andersen – University of Southern Denmark; Elisa Llurba Olive – Hospital Universitari Vall d'Hebron; Christopher Redman – University of Oxford; George Daskalakis – University of Athens; Maureen Macleod – University of Dundee; Baskaran Thilaganathan – St George's University of London; Javier Arenas Ramírez – University Hospital de Cabueñes; Jacques Massé – Laval University; Asma Khalil – St George's University of London; Francois Audibert – Université de Montréal; Per Minor Magnus – Norwegian Institute of Public Health; Anne Karen Jenum – University of Oslo; Ahmet Baschat – Johns Hopkins University School of Medicine; Akihide Ohkuchi – University School of Medicine, Shimotsuke-shi; Fionnuala M McAuliffe – University College Dublin; Jane West – University of Bristol; Lisa M Askie – University of Sydney; Fionnuala Mone – University College Dublin; Diane Farrar – Bradford Teaching Hospitals; Peter A Zimmerman – Päijät-Häme Central Hospital; Luc JM Smits – Maastricht University Medical Centre; Catherine Riddell – Better Outcomes Registry & Network (BORN); John C Kingdom – University of Toronto; Joris van de Post – Academisch Medisch Centrum; Sebastián E Illanes – University of the Andes; Claudia Holzman – Michigan State University; Sander MJ van Kuijk – Maastricht University Medical Centre; Lionel Carbillon – Assistance Publique-Hôpitaux de Paris Université; Pia M Villa – University of Helsinki and Helsinki University Hospital; Anne Eskild – University of Oslo; Lucy Chappell – King's College London; Federico Prefumo – University of Brescia; Luxmi Velauthar – Queen Mary University of London; Paul Seed – King's College London; Miriam van Oostwaard – IJsselland Hospital; Stefan Verlohren – Charité University Medicine; Lucilla Poston – King's College London; Enrico Ferrazzi – University of Milan; Christina A Vinter – University of Southern Denmark; Chie Nagata – National Center for Child Health and Development; Mark Brown – University of New South Wales; Karlijn C Vollebregt – Academisch Medisch Centrum; Satoru Takeda – Juntendo University; Josje Langenveld – Atrium Medisch Centrum Parkstad; Mariana Widmer – World Health Organization; Shigeru Saito – University of Toyama; Camilla Haavaldsen – Akershus University Hospital; Guillermo Carroli – Centro Rosarino De Estudios Perinatales; Jørn Olsen – Aarhus University; Hans Wolf – Academisch Medisch Centrum; Nelly Zavaleta – Instituto Nacional De Salud;

Inge Eisensee – Aarhus University; Patrizia Vergani – University of Milano-Bicocca; Pisake Lumbiganon – Khon Kaen University; Maria Makrides – South Australian Health and Medical Research Institute; Fabio Facchinetti – Università degli Studi di Modena e Reggio Emilia; Evan Sequeira – Aga Khan University; Marleen Temmerman – Aga Khan University; Robert Gibson – University of Adelaide; Sergio Ferrazzani – Università Cattolica del Sacro Cuore; Tiziana Frusca – Università degli Studi di Parma; Jane E Norman – University of Edinburgh; Ernesto A Figueiró-Filho – Mount Sinai Hospital; Olav Lapaire – Universitätsspital Basel; Hannele Laivuori – University of Helsinki and Helsinki University Hospital; Jacob A Lykke – Rigshospitalet; Agustin Conde-Agudelo – Eunice Kennedy Shriver National Institute of Child Health and Human Development; Alberto Galindo – Universidad Complutense de Madrid; Alfred Mbah – University of South Florida; Ana Pilar Betrán – World Health Organization; Ignacio Herraiz – Universidad Complutense de Madrid; Lill Trogstad – Norwegian Institute of Public Health; Gordon GS Smith – Cambridge University; Eric AP Steegers – University Hospital Nijmegen; Read Salim – HaEmek Medical Center; Tianhua Huang – North York General Hospital; Annemarijne Adank – Erasmus Medical Centre; Jun Zhang – National Institute of Child Health and Human Development; Wendy S Meschino – North York General Hospital; Joyce L Browne – University Medical Centre Utrecht; Rebecca E Allen – Queen Mary University of London; Fabricio Da Silva Costa – University of São Paulo; Kerstin Klipstein-Grobusch – University Medical Centre Utrecht; Caroline A Crowther – University of Adelaide; Jan Stener Jørgensen – Syddansk Universitet; Jean-Claude Forest – Centre hospitalier universitaire de Québec; Alice R Rumbold – University of Adelaide; Ben W Mol – Monash University; Yves Giguère – Laval University; Louise C Kenny – University of Liverpool; Wessel Ganzevoort – Academisch Medisch Centrum; Anthony O Odibo – University of South Florida; Jenny Myers – University of Manchester; SeonAe Yeo – University of North Carolina at Chapel Hill; Helena J Teede – Monash University and Monash Health; Francois Goffinet – Assistance publique – Hôpitaux de Paris; Lesley McCowan – University of Auckland; Eva Pajkrt – Academisch Medisch Centrum; Bassam G Haddad – Portland State University; Gustaaf Dekker – University of Adelaide; Emily C Kleinrouweler – Academisch Medisch Centrum; Édouard LeCarpentier – Centre Hospitalier Intercommunal Creteil; Claire T Roberts – University of Adelaide; Henk Groen – University Medical Center Groningen; Ragnhild Bergene Skråstad – St Olavs Hospital; Seppo Heinonen – University of Helsinki and Helsinki University Hospital; Kajantie Eero – University of Helsinki and Helsinki University Hospital; Dewi Anggrainin – Lambung Mangkurat University; Athena Souka – University of Athens; Jose Cecatti – University of Campinas; Arri Coomarasamy – University of Birmingham; Athanasios Pilalis – University of Athens; Renato T Souza – University of Campinas; Lee Ann Hawkins – University of Calgary; Rinat Gabbay-Benziv – Hillel Yaffe Medical Center; Francesc Figueras – University of Barcelona; Francesca Crovetto – University of Barcelona; Marleen van Gelder – Radboud University Medical Center; Line Sletner – Akershus University Hospital and University of Oslo.

Contributions of authors

John Allotey (https://orcid.org/0000-0003-4134-6246) (Lecturer, Women's Health) developed the protocol, undertook the literature searches, study selection, drafted the manuscript, acquired Individual Participant Data, led the project, mapped the IPD variables and cleaned and quality checked the data.

Lucinda Archer (https://orcid.org/0000-0003-2504-2613) (Lecturer, Biostatistics) undertook the literature searches, study selection, drafted the manuscript, acquired Individual Participant Data, led the project and conducted the data analysis.

Dyuti Coomar (https://orcid.org/0000-0001-6229-0830) (Research Fellow, Biostatistics) conducted the economic evaluation.

Kym IE Snell (https://orcid.org/0000-0001-9373-6591) (Lecturer, Biostatistics) undertook the literature searches, study selection, drafted the manuscript, acquired Individual Participant Data, led the project and conducted the data analysis.

Melanie Smuk (https://orcid.org/0000-0002-1594-1458) (Senior Lecturer, Medical Statistics) undertook the literature searches, study selection, drafted the manuscript, acquired Individual Participant Data, led the project, mapped the IPD variables, cleaned and quality checked the data and harmonised the data.

Lucy Oakey (https://orcid.org/0000-0001-7664-5428) (Research Manager, Women's Health) provided patient and public input and wrote the Plain language summary.

Sadia Haqnawaz (https://orcid.org/0000-0001-6777-8712) (PPI) provided patient and public input and wrote the Plain language summary.

Ana Pilar Betrán (https://orcid.org/0000-0002-5631-5883) (Medical Officer, Reproductive Health) developed the protocol.

Lucy C Chappell (https://orcid.org/0000-0001-6219-3379) (Professor, Obstetrics) developed the protocol.

Wessel Ganzevoort (https://orcid.org/0000-0002-7243-2115) (Specialist, Obstetrician and Gynaecologist) developed the protocol.

Sanne Gordijn (https://orcid.org/0000-0003-3915-8609) (Specialist, Obstetrician and Gynaecologist) developed the protocol.

Asma Khalil (https://orcid.org/0000-0003-2802-7670) (Professor, Obstetrics and Maternal-Fetal Medicine) developed the protocol.

Ben W Mol (https://orcid.org/0000-0001-8337-550X) (Professor, Obstetrics and Gynaecology) developed the protocol.

Rachel K Morris (https://orcid.org/0000-0003-1247-429X) (Professor, Obstetrics) developed the protocol.

Jenny Myers (https://orcid.org/0000-0003-0913-2096) (Professor, Maternal & Fetal Health) developed the protocol.

Aris T Papageorghiou (https://orcid.org/0000-0001-8143-2232) (Professor, Obstetrics and Maternal-Fetal Medicine) developed the protocol.

Basky Thilaganathan (https://orcid.org/0000-0002-5531-4301) (Professor, Fetal Medicine) developed the protocol.

Fabricio Da Silva Costa (https://orcid.org/0000-0002-0765-7780) (Adjunct Clinical A/Prof, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Fabio Facchinetti (https://orcid.org/0000-0003-4694-9564) (Director, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Arri Coomarasamy (https://orcid.org/0000-0002-3261-9807) (Professor, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Akihide Ohkuchi (https://orcid.org/0000-0002-8861-1572) (Professor, Obstetrics and Gynecology) provided input into the drafting of the initial manuscript.

Anne Eskild (https://orcid.org/0000-0002-2756-1583) (Professor, Clinic Medicine) provided input into the drafting of the initial manuscript.

Javier Arenas Ramírez (https://orcid.org/0000-0003-2291-720X) (Researcher, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Alberto Galindo (https://orcid.org/0000-0002-1334-4879) (Researcher, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Ignacio Herraiz (https://orcid.org/0000-0001-6807-4944) (Researcher, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Federico Prefumo (https://orcid.org/0000-0001-7793-714X) (Specialist, Gynaecology and Obstetrics) provided input into the drafting of the initial manuscript.

Shigeru Saito (https://orcid.org/0000-0002-8940-3708) (Professor, Obstetrics and Gynecology) provided input into the drafting of the initial manuscript.

Line Sletner (https://orcid.org/0000-0002-5085-7366) (Postdoc researcher, Metabolic Health) provided input into the drafting of the initial manuscript.

Jose Guilherm Cecatti (https://orcid.org/0000-0003-1285-8445) (Professor, Obstetrics and Gynecology) provided input into the drafting of the initial manuscript.

Rinat Gabbay-Benziv (https://orcid.org/0000-0001-6887-6314) (Professor, Obstetrics and Gynecology) provided input into the drafting of the initial manuscript.

Francois Goffinet (https://orcid.org/0000-0001-9643-0299) (Professor, Gynecology and Obstetrics) provided input into the drafting of the initial manuscript.

Ahmet A Baschat (https://orcid.org/0000-0003-1927-2084) (Professor, Gynaecology and Obstetrics) provided input into the drafting of the initial manuscript.

Renato T Souza (https://orcid.org/0000-0002-9075-9269) (Research Fellow, Maternal and Perinatal Health) provided input into the drafting of the initial manuscript.

Fionnuala Mone (https://orcid.org/0000-0002-0718-7547) (Clinical Lecturer, Maternal Fetal Medicine) provided input into the drafting of the initial manuscript.

Diane Farrar (https://orcid.org/0000-0002-5625-761X) (NIHR postdoctoral research fellow, Maternal Health) provided input into the drafting of the initial manuscript.

Seppo Heinonen (https://orcid.org/0000-0001-5949-0874) (Professor, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Kjell Å Salvesen (https://orcid.org/0000-0002-1788-4063) (Professor, Clinical and Molecular Medicine) provided input into the drafting of the initial manuscript.

Luc JM Smits (https://orcid.org/0000-0003-0785-1345) (Professor, Epidemiology) provided input into the drafting of the initial manuscript.

Sohinee Bhattacharya (https://orcid.org/0000-0002-2358-5860) (Senior Lecturer, Medicine) provided input into the drafting of the initial manuscript.

Chie Nagata (https://orcid.org/0000-0003-4897-2119) (Chief of Clinical Research, Maternal and Child Health) provided input into the drafting of the initial manuscript.

Satoru Takeda (https://orcid.org/0000-0001-9547-1435) (Professor, Obstetrics and Gynaecology) provided input into the drafting of the initial manuscript.

Marleen MHJ van Gelder (https://orcid.org/0000-0003-4853-4434) (Assistant Professor, Perinatal Pharmacoepidemiology) provided input into the drafting of the initial manuscript.

Dewi Anggraini (https://orcid.org/0000-0001-7507-2445) (Consultant, Pediatrics) provided input into the drafting of the initial manuscript.

SeonAe Yeo (https://orcid.org/0000-0002-0721-0997) (Professor, Nursing) provided input into the drafting of the initial manuscript.

Jane West (https://orcid.org/0000-0002-5770-8363) (Professor, Public Health) provided input into the drafting of the initial manuscript.

Javier Zamora (https://orcid.org/0000-0003-4901-588X) (Professor, Biostatistics) developed the protocol.

Hema Mistry (https://orcid.org/0000-0002-5023-1160) (Associate Professor, Clinical Trials and Health Economics) developed the protocol and conducted the economic evaluation.

Richard D Riley (https://orcid.org/0000-0001-8699-0735) (Professor, Biostatistics) developed the protocol and conducted the data analysis.

Shakila Thangaratinam (https://orcid.org/0000-0002-4254-460X) (Professor, Maternal and Perinatal Health) developed the protocol, undertook the literature searches, study selection, drafted the manuscript, acquired Individual Participant Data and led the project.

All authors critically appraised and provided feedback and input into the drafts and final version of the report.

Ethics statement

Ethics approval was not required because the IPD meta-analysis involved secondary analysis of existing anonymised data.

Patient and public involvement

PPI members provided input to the running of the project via participation in the steering committee and project management groups. Katie's team members which include mothers, pregnant women, carers and family members with an interest in improving the quality of research within women's health, contributed to the fine-tuning of the primary outcomes of the project proposal, by providing feedback on what they would consider to be an important outcome. Development of the prediction model also took into account input from service users on the acceptability of predictors included in the model. A lay member of the Hilda's group (a public advisory group for women's health research) wrote the *Plain language summary* of the report. Dissemination of findings will be done in collaboration with Katie's team, the Hilda's, Sands charity and other interested groups.

Data-sharing statement

All data requests should be submitted to the corresponding author for consideration. Access to available anonymised data may be granted following review and appropriate agreements being in place.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Health Technology Assessment programme (NIHR award ref: 17/148/07) and is published in full in *Health Technology Assessment*; Vol. 28, No. 47. See the NIHR Funding and Awards website for further award information.

References

1. Mamelle N, Cochet V, Claris O. Definition of fetal growth restriction according to constitutional growth potential. *Biol Neonate* 2001;**80**(4):277–85.
2. Divon MY, Haglund B, Nisell H, Otterblad PO, Westgren M. Fetal and neonatal mortality in the postterm pregnancy: the impact of gestational age and fetal growth restriction. *Am J Obstet Gynecol* 1998;**178**(4):726–31.
3. Barker DJ, Osmond C, Golding J, Kuh D, Wadsworth ME. Growth in utero, blood pressure in childhood and adult life, and mortality from cardiovascular disease. *BMJ* 1989;**298**(6673):564–7.
4. von Beckerath AK, Kollmann M, Rotky-Fast C, Karpf E, Lang U, Klaritsch P. Perinatal complications and long-term neurodevelopmental outcome of infants with intrauterine growth restriction. *Am J Obstet Gynecol* 2013;**208**(2):130 e1–6.
5. Wilcox AJ, Cortese M, McConnaughey DR, Moster D, Basso O. The limits of small-for-gestational-age as a high-risk category. *Eur J Epidemiol* 2021;**36**(10):985–91.
6. Gardosi J, Madurasinghe V, Williams M, Malik A, Francis A. Maternal and fetal risk factors for stillbirth: population based study. *BMJ* 2013;**346**:f108.
7. Vayssiere C, Sentilhes L, Ego A, Bernard C, Cambourieu D, Flamant C, *et al.* Fetal growth restriction and intra-uterine growth restriction: guidelines for clinical practice from the French College of Gynaecologists and Obstetricians. *Eur J Obstet Gynecol Reprod Biol* 2015;**193**:10–8.
8. *Saving Babies' Lives Version Two: A Care Bundle for Reducing Perinatal Mortality*. URL: www.england.nhs.uk/wp-content/uploads/2019/03/Saving-Babies-Lives-Care-Bundle-Version-Two-Updated-Final-Version.pdf (accessed 15 October 2020).
9. Dall'Asta A, Brunelli V, Prefumo F, Frusca T, Lees CC. Early onset fetal growth restriction. *Matern Health Neonatol Perinatol* 2017;**3**:2.
10. Hepburn M, Rosenberg K. An audit of the detection and management of small-for-gestational age babies. *Br J Obstet Gynaecol* 1986;**93**(3):212–6.
11. Royal College of Obstetricians and Gynaecologists. *The Investigation and Management of the Small–for–Gestational–Age Fetus – RCOG Green-top Guideline No. 31: 2nd Edition. February 2013. Minor revisions – January 2014*. URL: www.rcog.org.uk/globalassets/documents/guidelines/gtg_31.pdf (accessed 29 June 2022).
12. American College of Obstetricians, Gynecologists Committee on Practice Bulletins – Obstetrics SfM-FMPC. Fetal growth restriction: ACOG practice bulletin, number 227. *Obstet Gynecol* 2021;**137**(2):e16–28.
13. Lausman A, Kingdom J, Gagnon R, *et al.* Intrauterine growth restriction: screening, diagnosis, and management. *J Obstet Gynaecol Can* 2013;**35**(8):741–8.
14. The Royal Australian and New Zealand College of Obstetricians and Gynaecologists. *Screening in Early Pregnancy for Adverse Perinatal Outcomes. Prenatal Screening for Adverse Pregnancy Outcomes. C-Obs 61 (July 2015)*. URL: https://ranzcog.edu.au/wp-content/uploads/2022/05/Screening-in-Early-Pregnancy-for-Adverse-Perinatal-Outcomes.pdf (accessed 29 June 2022).
15. Lees CC, Stampalija T, Baschat A, *et al.* ISUOG Practice Guidelines: diagnosis and management of small-for-gestational-age fetus and fetal growth restriction. *Ultrasound Obstet Gynecol* 2020;**56**(2):298–312.

16. Melamed N, Baschat A, Yinon Y, *et al.* FIGO (International Federation of Gynecology and Obstetrics) initiative on fetal growth: best practice advice for screening, diagnosis, and management of fetal growth restriction. *Int J Gynaecol Obstet* 2021;**152**(Suppl. 1):3–57.

17. Bricker L, Medley N, Pratt JJ. Routine ultrasound in late pregnancy (after 24 weeks' gestation). *Cochrane Database Syst Rev* 2015;(6):CD001451.

18. Sovio U, White IR, Dacey A, Pasupathy D, Smith GCS. Screening for fetal growth restriction with universal third trimester ultrasonography in nulliparous women in the Pregnancy Outcome Prediction (POP) study: a prospective cohort study. *Lancet* 2015;**386**(10008):2089–97.

19. Monier I, Blondel B, Ego A, Kaminiski M, Goffinet F, Zeitlin J. Poor effectiveness of antenatal detection of fetal growth restriction and consequences for obstetric management and neonatal outcomes: a French national study. *BJOG* 2015;**122**(4):518–27.

20. Henrichs J, Verfaille V, Jellema P, *et al.* Effectiveness of routine third trimester ultrasonography to reduce adverse perinatal outcomes in low risk pregnancy (the IRIS study): nationwide, pragmatic, multicentre, stepped wedge cluster randomised trial. *BMJ* 2019;**367**:l5517.

21. National Institute for Health and Care Excellence (NICE). *Antenatal Care for Uncomplicated Pregnancies*. Clinical Guideline [CG62]. London: NICE; 2008. URL: https://www.nice.org.uk/guidance/cg62 (accessed 21 March 2018).

22. Morriss RK. *Prediction and Prevention of Fetal Growth Restriction and Compromise of Fetal Wellbeing. Systematic Reviews and Meta-Analyses with Model Based Economic Evaluation*. A thesis submitted to the University of Birmingham for the degree of Doctor of Philosophy. College of Medical and Dental Sciences, The University of Birmingham; September 2010, Volume I. URL: http://etheses.bham.ac.uk/1319/1/Morris11PhD.pdf (accessed 5 May 2018).

23. Morris RK, Cnossen JS, Langejans M, *et al.* Serum screening with Down's syndrome markers to predict pre-eclampsia and small for gestational age: systematic review and meta-analysis. *BMC Pregnancy Childbirth* 2008;**8**:33.

24. Morris RK, Khan KS, Coomarasamy A, Robson SC, Kleijnen J. The value of predicting restriction of fetal growth and compromise of its wellbeing: systematic quantitative overviews (meta-analysis) of test accuracy literature. *BMC Pregnancy Childbirth* 2007;**7**:3.

25. Morris RK, Bilagi A, Devani P, Kilby MD. Association of serum PAPP-A levels in first trimester with small for gestational age and adverse pregnancy outcomes: systematic review and meta-analysis. *Prenat Diagn* 2017;**37**(3):253–65.

26. Kleinrouweler CE, Cheong-See FM, Collins GS, *et al.* Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016 Jan;**214**(1):79–90.e36. https://doi.org/10.1016/j.ajog.2015.06.013. *Epub* 2015 Jun 10. PMID: 26070707.

27. Moons KG, Altman DG, Reitsma JB, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015;**162**(1):W1–73.

28. World Health Organization (WHO). *Children's Environmental Health Indicators. Intrauterine Growth Retardation in Newborn Children*. URL: www.who.int/ceh/indicators/iugrnewborn.pdf (accessed 5 March 2018).

29. Zhang X, Platt RW, Cnattingius S, Joseph KS, Kramer MS. The use of customised versus population-based birthweight standards in predicting perinatal mortality. *BJOG* 2007;**114**(4):474–7.

30. McCowan LM, Harding JE, Stewart AW. Customized birthweight centiles predict SGA pregnancies with perinatal morbidity. *BJOG* 2005;**112**(8):1026–33.

31. Riley RD, Ensor J, Snell KI. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016;**353**:i3140. https://doi.org/10.1136/bmj.i3140

32. Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med* 2013;**32**(18):3158–80.
33. Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG; Cochrane IPDM-aMg. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. *PLOS Med* 2015;**12**(10):e1001886.
34. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;**340**:c221.
35. Riley RD, Tierney J, Stewart LA, editors. *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chichester: Wiley; 2021.
36. Allotey J, Snell KI, Smuk M, *et al.* Validation and development of models using clinical, biochemical and ultrasound markers for predicting pre-eclampsia: an individual participant data meta-analysis. *Health Technol Assess* 2020;**24**(72):1–252.
37. Gordijn SJ, Beune IM, Thilaganathan B, *et al.* Consensus definition of fetal growth restriction: a Delphi procedure. *Ultrasound Obstet Gynecol* 2016;**48**(3):333–9.
38. Sovio U, Smith GCS. The effect of customization and use of a fetal growth standard on the association between birthweight percentile and adverse perinatal outcome. *Am J Obstet Gynecol* 2018;**218**(2S):S738–S44.
39. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;**338**:b605.
40. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;**338**:b375.
41. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ* 2009;**338**:b604.
42. Riley RD, van der Windt D, Croft P, *et al.*, editors. *Prognosis Research in Healthcare: Concepts, Methods and Impact*. Oxford, UK: Oxford University Press; 2019.
43. Stewart LA, Clarke M, Rovers M, *et al.* Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA* 2015;**313**(16):1657–65.
44. Allotey J, Snell KIE, Chan C, *et al.* External validation, update and development of prediction models for pre-eclampsia using an Individual Participant Data (IPD) meta-analysis: the International Prediction of Pregnancy Complication Network (IPPIC pre-eclampsia) protocol. *Diagn Progn Res* 2017;**1**:16.
45. Snell KIE, Allotey J, Smuk M, *et al.* External validation of prognostic models predicting pre-eclampsia: individual participant data meta-analysis. *BMC Med* 2020;**18**(1):302.
46. Townsend R, Sileo F, Stocker L, *et al.* Variation in outcome reporting in randomized controlled trials of interventions for prevention and treatment of fetal growth restriction. *Ultrasound Obstet Gynecol* 2019;**53**(5):598–608.
47. Wolff RF, Moons KGM, Riley RD, *et al.* PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019;**170**(1):51–8.
48. Cheong-See F, Allotey J, Marlin N, *et al.* Prediction models in obstetrics: understanding the treatment paradox and potential solutions to the threat it poses. *BJOG* 2016;**123**(7):1060–4.
49. Poon LC, Tan MY, Yerlikaya G, Syngelaki A, Nicolaides KH. Birth weight in live births and stillbirths. *Ultrasound Obstet Gynecol* 2016;**48**(5):602–6.

50. Papageorghiou AT, Ohuma EO, Altman DG, *et al.* International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *Lancet* 2014;**384**(9946):869–79.

51. de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martines J. The WHO Multicentre Growth Reference Study: planning, study design, and methodology. *Food Nutr Bull* 2004;**25**(1 Suppl.):S15–26.

52. Heazell AEP, Hayes DJL, Whitworth M, Takwoingi Y, Bayliss SE, Davenport C. Diagnostic accuracy of biochemical tests of placental function versus ultrasound assessment of fetal size for stillbirth and small-for-gestational-age infants. *Cochrane Database Syst Rev* 2019 May 14;**5**(5):CD012245. https://doi.org/10.1002/14651858.CD012245.pub2. PMID: 31087568; PMCID: PMC6515632.

53. Riley RD, Debray TPA, Collins GS, *et al.* Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med* 2022 Mar 30;**41**(7):1280–1295. https://doi.org/10.1002/sim.9275. *Epub* 2021 Dec 16. PMID: 34915593.

54. 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**(2):214–26.

55. Archer L, Snell KIE, Ensor J, Hudda MT, Collins GS, Riley RD. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Stat Med* 2021;**40**(1):133–46.

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

57. Roberts LA, Ling HZ, Poon LC, Nicolaides KH, Kametas NA. Maternal hemodynamics, fetal biometry and Doppler indices in pregnancies followed up for suspected fetal growth restriction. *Ultrasound Obstet Gynecol* 2018;**52**(4):507–14.

58. Riley RD, Snell KIE, Ensor J, *et al.* Minimum sample size for developing a multivariable prediction model: Part I – Continuous outcomes. *Stat Med* 2019;**38**(7):1262–75.

59. 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**(1):58–65.

60. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 2011;**30**(4):377–99.

61. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd edn. New York: John Wiley; 2002.

62. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York, NY: Springer; 2009.

63. Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**(1):128–38.

64. Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. *Biom J* 2015;**57**(4):614–32.

65. Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med* 2017;**36**(5):855–75.

66. Debray TP, Damen JA, Snell KI, *et al.* A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017;**356**:i6460.

67. Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;**342**:d549.

68. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the *C*-statistic and calibration measures? *Stat Methods Med Res* 2018;**27**(11):3505–22.
69. Rover C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol* 2015;**15**:99.
70. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;**172**(1):137–59.
71. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006;**26**(6):565–74.
72. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016;**352**:i6.
73. Jolani S, Debray TP, Koffijberg H, van Buuren S, Moons KG. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med* 2015;**34**(11):1841–63.
74. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. *Stat Med* 1990;**9**(11):1303–25.
75. Fraser A, Macdonald-Wallis C, Tilling K, *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013;**42**(1):97–110.
76. *Ontario's Better Outcomes Registry & Network (BORN)*. URL: www.bornontario.ca/en/about-born/ (accessed 14 March 2019).
77. *Japan Society of Obstetrics and Gynecology*. URL: www.jsog.or.jp/modules/en/index.php?content_id=1 (accessed 14 March 2019).
78. Carter J, Seed PT, Tribe RM, *et al.* Saliva progesterone for prediction of spontaneous preterm birth: The POPPY study. Pregnancy Outcome Poster Abstracts. *BJOG* 2017;**124**:122–54.
79. Al-Amin A, Rolnik DL, Black C, *et al.* Accuracy of second trimester prediction of preterm preeclampsia by three different screening algorithms. *Aust N Z J Obstet Gynaecol* 2018;**58**(2):192–6.
80. Allen RE, Zamora J, Arroyo-Manzano D, *et al.* External validation of preexisting first trimester preeclampsia prediction models. *Eur J Obstet Gynecol Reprod Biol* 2017;**217**:119–25.
81. Andersen LB, Dechend R, Jorgensen JS, *et al.* Prediction of preeclampsia with angiogenic biomarkers. Results from the prospective Odense Child Cohort. *Hypertens Pregnancy* 2016;**35**(3):405–19.
82. Antsaklis A, Daskalakis G, Tzortzis E, Michalas S. The effect of gestational age and placental location on the prediction of pre-eclampsia by uterine artery Doppler velocimetry in low-risk nulliparous women. *Ultrasound Obstet Gynecol* 2000;**16**(7):635–9.
83. Arenas J, Fernández-iñarea J, Rodríguez-mon C, Duplá B, Díez E, González-garcía A. Cribado con doppler de las arterias uterinas para la predicción de complicaciones de la gestación. *Clínica e Investigación en Ginecología y Obstetricia* 2003;**30**(6):178–84.
84. Askie LM, Duley L, Henderson-Smart DJ, Stewart LA. Antiplatelet agents for prevention of pre-eclampsia: a meta-analysis of individual patient data. *Lancet* 2007;**369**(9575):1791–8.
85. Audibert F, Boucoiran I, An N, *et al.* Screening for preeclampsia using first-trimester serum markers and uterine artery Doppler in nulliparous women. *Am J Obstet Gynecol* 2010;**203**(4):383 e1–8.
86. Ayorinde AA, Wilde K, Lemon J, Campbell D, Bhattacharya S. Data resource profile: The Aberdeen Maternity and Neonatal Databank (AMND). *Int J Epidemiol* 2016;**45**(2):389–94.

87. Baschat AA, Magder LS, Doyle LE, Atlas RO, Jenkins CB, Blitzer MG. Prediction of preeclampsia utilizing the first trimester screening examination. *Am J Obstet Gynecol* 2014;**211**(5):514.e1–7.

88. Brown MA, Mackenzie C, Dunsmuir W, *et al.* Can we predict recurrence of pre-eclampsia or gestational hypertension? *BJOG* 2007;**114**(8):984–93.

89. Cameroni I, Roncaglia N, Crippa I, *et al.* P32.05: uterine artery Doppler in a risk population: what's its role in the prediction of severe pregnancy complications? *Ultrasound Obstet Gynecol* 2008;**32**(3):421–2.

90. Caradeux J, Serra R, Nien JK, *et al.* First trimester prediction of early onset preeclampsia using demographic, clinical, and sonographic data: a cohort study. *Prenat Diagn* 2013;**33**(8):732–6.

91. Carbillon L. The imbalance of circulating angiogenic/antiangiogenic factors is mild or absent in obese women destined to develop preeclampsia. *Hypertens Pregnancy* 2014;**33**(4):524.

92. Caritis S, Sibai B, Hauth J, *et al.* Low-dose aspirin to prevent preeclampsia in women at high risk. National Institute of Child Health and Human Development Network of Maternal-Fetal Medicine Units. *N Engl J Med* 1998;**338**(11):701–5.

93. Chappell LC, Seed PT, Briley AL, *et al.* Effect of antioxidants on the occurrence of pre-eclampsia in women at increased risk: a randomised trial. *Lancet* 1999;**354**(9181):810–6.

94. Chiswick C, Reynolds RM, Denison F, *et al.* Effect of metformin on maternal and fetal outcomes in obese pregnant women (EMPOWaR): a randomised, double-blind, placebo-controlled trial. *Lancet Diabetes Endocrinol* 2015;**3**(10):778–86.

95. Conserva V, Muggiasca M, Arrigoni L, Mantegazza V, Rossi E, Ferrazzi E. Recurrence and severity of abnormal pregnancy outcome in patients treated by low-molecular-weight heparin: a prospective pilot study. *J Matern Fetal Neonatal Med* 2012;**25**(8):1467–73.

96. Facchinetti F, Marozio L, Frusca T, *et al.* Maternal thrombophilia and the risk of recurrence of preeclampsia. *Am J Obstet Gynecol* 2009;**200**(1):46.e1–5.

97. Figueiró-Filho EA, Oliveira VM, Coelho LR, Breda I. Marcadores séricos de trombofilias hereditárias e anticorpos antifosfolípides em gestantes com antecedentes de pré-eclâmpsia grave. *Revista Brasileira de Ginecologia e Obstetrícia* 2012;**34**(1):40–6.

98. Giguere Y, Masse J, Theriault S, *et al.* Screening for pre-eclampsia early in pregnancy: performance of a multivariable model combining clinical characteristics and biochemical markers. *BJOG* 2015;**122**(3):402–10.

99. Girchenko P, Lahti M, Tuovinen S, *et al.* Cohort Profile: prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) study. *Int J Epidemiol* 2017;**46**(5):1380–1g.

100. Goetzinger KR, Singla A, Gerkowicz S, Dicke JM, Gray DL, Odibo AO. Predicting the risk of pre-eclampsia between 11 and 13 weeks' gestation by combining maternal characteristics and serum analytes, PAPP-A and free beta-hCG. *Prenat Diagn* 2010;**30**(12-13):1138–42.

101. Goffinet F, Aboulker D, Paris-Llado J, *et al.* Screening with a uterine Doppler in low risk pregnant women followed by low dose aspirin in women with abnormal results: a multicenter randomised controlled trial. *BJOG* 2001;**108**(5):510–8.

102. Gurgel Alves JA, Praciano de Sousa PC, Bezerra Maia EHMS, Kane SC, da Silva Costa F. First-trimester maternal ophthalmic artery Doppler analysis for prediction of pre-eclampsia. *Ultrasound Obstet Gynecol* 2014;**44**(4):411–8.

103. Holzman C, Bullen B, Fisher R, Paneth N, Reuss L; Prematurity Study G. Pregnancy outcomes and community health: the POUCH study of preterm delivery. *Paediatr Perinat Epidemiol* 2001;**15**(Suppl. 2):136–58.

104. Huang T, Hoffman B, Meschino W, Kingdom J, Okun N. Prediction of adverse pregnancy outcomes by combinations of first and second trimester biochemistry markers used in the routine prenatal screening of Down syndrome. *Prenat Diagn* 2010;**30**(5):471–7.

105. Jaaskelainen T, Heinonen S, Hamalainen E, *et al.* Angiogenic profile in the Finnish Genetics of Pre-Eclampsia Consortium (FINNPEC) cohort. *Pregnancy Hypertens* 2018;**14**:252–9.

106. Jaddoe VW, van Duijn CM, Franco OH, *et al.* The Generation R Study: design and cohort update 2012. *Eur J Epidemiol* 2012;**27**(9):739–56.

107. Jenum AK, Sletner L, Voldner N, *et al.* The STORK Groruddalen research programme: a population-based cohort study of gestational diabetes, physical activity, and obesity in pregnancy in a multiethnic population. Rationale, methods, study population, and participation rates. *Scand J Public Health* 2010;**38**(5 Suppl.):60–70.

108. Olsen J, Melbye M, Olsen SF, *et al.* The Danish National Birth Cohort – its background, structure and aim. *Scand J Public Health* 2001;**29**:300–7.

109. Khan F, Belch JJ, MacLeod M, Mires G. Changes in endothelial function precede the clinical disease in women in whom preeclampsia develops. *Hypertension* 2005;**46**(5):1123–8.

110. Langenveld J, Buttinger A, van der Post J, Wolf H, Mol BW, Ganzevoort W. Recurrence risk and prediction of a delivery under 34 weeks of gestation after a history of a severe hypertensive disorder. *BJOG* 2011;**118**(5):589–95.

111. Lecarpentier E, Tsatsaris V, Goffinet F, Cabrol D, Sibai B, Haddad B. Risk factors of superimposed preeclampsia in women with essential chronic hypertension treated before pregnancy. *PLOS ONE* 2013;**8**(5):e62140.

112. Llurba E, Carreras E, Gratacos E, *et al.* Maternal history and uterine artery Doppler in the assessment of risk for development of early- and late-onset preeclampsia and intrauterine growth restriction. *Obstet Gynecol Int* 2009;**2009**:275613.

113. Lykke JA, Paidas MJ, Langhoff-Roos J. Recurring complications in second pregnancy. *Obstet Gynecol* 2009;**113**(6):1217–24.

114. Magnus P, Irgens LM, Haug K, *et al.* Cohort profile: the Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol* 2006;**35**(5):1146–50.

115. Makrides M, Gibson RA, McPhee AJ, *et al.* Effect of DHA supplementation during pregnancy on maternal depression and neurodevelopment of young children: a randomized controlled trial. *JAMA* 2010;**304**(15):1675–83.

116. Massé J, Forest JC, Moutquin JM, Marcoux S, Brideau NA, Bélanger M. A prospective study of several potential biologic markers for early prediction of the development of preeclampsia. *Am J Obstet Gynecol* 1993;**3**(169):501–8.

117. Mbah AK, Sharma PP, Alio AP, Fombo DW, Bruder K, Salihu HM. Previous cesarean section, gestational age at first delivery and subsequent risk of pre-eclampsia in obese mothers. *Arch Gynecol Obstet* 2012;**285**(5):1375–81.

118. Mone F, Mulcahy C, McParland P, *et al.* An open-label randomized-controlled trial of low dose aspirin with an early screening test for pre-eclampsia and growth restriction (TEST): trial protocol. *Contemp Clin Trials* 2016;**49**:143–8.

119. North RA, McCowan LM, Dekker GA, *et al.* Clinical risk prediction for pre-eclampsia in nulliparous women: development of model in international prospective cohort. *BMJ* 2011;**342**:d1875.

120. Odibo AO, Zhong Y, Goetzinger KR, *et al.* First-trimester placental protein 13, PAPP-A, uterine artery Doppler and maternal characteristics in the prediction of pre-eclampsia. *Placenta* 2011;**32**(8):598–602.

121. Ohkuchi A, Minakami H, Sato I, Mori H, Nakano T, Tateno M. Predicting the risk of pre-eclampsia and a small-for-gestational-age infant by quantitative assessment of the diastolic notch in uterine artery flow velocity waveforms in unselected women. *Ultrasound Obstet Gynecol* 2000;**16**(2):171–8.

122. Poston L, Bell R, Croker H, *et al.* Effect of a behavioural intervention in obese pregnant women (the UPBEAT study): a multicentre, randomised controlled trial. *Lancet Diabetes Endocrinol* 2015;**3**(10):767–77.

123. Poston L, Briley AL, Seed PT, Kelly FJ, Shennan AH. Vitamin C and vitamin E in pregnant women at risk for pre-eclampsia (VIP trial): randomised placebo-controlled trial. *Lancet* 2006;**367**(9517):1145–54.

124. Prefumo F, Fratelli N, Ganapathy R, Bhide A, Frusca T, Thilaganathan B. First trimester uterine artery Doppler in women with previous pre-eclampsia. *Acta Obstet Gynecol Scand* 2008;**87**(12):1271–5.

125. Rang S, van Montfrans GA, Wolf H. Serial hemodynamic measurement in normal pregnancy, preeclampsia, and intrauterine growth restriction. *Am J Obstet Gynecol* 2008;**198**(5):519.e1–9.

126. Rocha RS, Alves JAG, Maia EHMSB, *et al.* Simple approach based on maternal characteristics and mean arterial pressure for the prediction of preeclampsia in the first trimester of pregnancy. *J Perinat Med* 2017;**45**(7):843–9.

127. Rocha RS, Gurgel Alves JA, Bezerra Maia EHMS, *et al.* Comparison of three algorithms for prediction preeclampsia in the first trimester of pregnancy. *Pregnancy Hypertens* 2017;**10**:113–7.

128. Rumbold AR, Crowther CA, Haslam RR, Dekker GA, Robinson JS, Group AS. Vitamins C and E and the risks of preeclampsia and perinatal complications. *N Engl J Med* 2006;**354**(17):1796–806.

129. Salim R, Czarnowicki T, Nachum Z, Shalev E. The impact of close surveillance on pregnancy outcome among women with a prior history of antepartum complications attributed to thrombosis: a cohort study. *Reprod Biol Endocrinol* 2008;**6**:55.

130. Savitri AI, Zuithoff P, Browne JL, *et al.* Does pre-pregnancy BMI determine blood pressure during pregnancy? A prospective cohort study. *BMJ Open* 2016;**6**(8):e011626.

131. Sergio F, Maria Clara D, Gabriella F, *et al.* Prophylaxis of recurrent preeclampsia: low-molecular-weight heparin plus low-dose aspirin versus low-dose aspirin alone. *Hypertens Pregnancy* 2006;**25**(2):115–27.

132. Sibai BM, Caritis SN, Thom E, *et al.* Prevention of preeclampsia with low-dose aspirin in healthy, nulliparous pregnant women. The National Institute of Child Health and Human Development Network of Maternal-Fetal Medicine Units. *N Engl J Med* 1993;**329**(17):1213–8.

133. Skrastad RB, Hov GG, Blaas HG, Romundstad PR, Salvesen KA. Risk assessment for pre-eclampsia in nulliparous women at 11–13 weeks gestational age: prospective evaluation of two algorithms. *BJOG* 2015;**122**(13):1781–8.

134. Staff AC, Braekke K, Harsem NK, Lyberg T, Holthe MR. Circulating concentrations of sFlt1 (soluble fms-like tyrosine kinase 1) in fetal and maternal serum during pre-eclampsia. *Eur J Obstet Gynecol Reprod Biol* 2005;**122**(1):33–9.

135. Stirrup OT, Khalil A, D'Antonio F, Thilaganathan B; Southwest Thames Obstetric Research Collaborative (STORK). Fetal growth reference ranges in twin pregnancy: analysis of the Southwest Thames Obstetric Research Collaborative (STORK) multiple pregnancy cohort. *Ultrasound Obstet Gynecol* 2015;**45**(3):301–7.

136. Trogstad L, Skrondal A, Stoltenberg C, Magnus P, Nesheim BI, Eskild A. Recurrence risk of preeclampsia in twin and singleton pregnancies. *Am J Med Genet A* 2004;**126A**(1):41–5.

137. Van Der Linden EL, Browne JL, Vissers KM, *et al.* Maternal body mass index and adverse pregnancy outcomes: a Ghanaian cohort study. *Obesity (Silver Spring)* 2016;**24**(1):215–22.

138. van Kuijk SM, Delahaije DH, Dirksen CD, *et al.* External validation of a model for periconceptional prediction of recurrent early-onset preeclampsia. *Hypertens Pregnancy* 2014;**33**(3):265–76.

139. van Kuijk SM, Nijdam ME, Janssen KJ, *et al.* A model for preconceptional prediction of recurrent early-onset preeclampsia: derivation and internal validation. *Reprod Sci (Thousand Oaks, Calif)* 2011;**18**(11):1154–9.

140. van Oostwaard MF, Langenveld J, Bijloo R, *et al.* Prediction of recurrence of hypertensive disorders of pregnancy between 34 and 37 weeks of gestation: a retrospective cohort study. *BJOG* 2012;**119**(7):840–7.

141. Van Oostwaard MF, Langenveld J, Schuit E, *et al.* Prediction of recurrence of hypertensive disorders of pregnancy in the term period, a retrospective cohort study. *Pregnancy Hypertens* 2014;**4**(3):194–202.

142. Vatten LJ, Eskild A, Nilsen TI, Jeansson S, Jenum PA, Staff AC. Changes in circulating level of angiogenic factors from the first to second trimester as predictors of preeclampsia. *Am J Obstet Gynecol* 2007;**196**(3):239.e1–6.

143. Verlohren S, Galindo A, Schlembach D, *et al.* An automated method for the determination of the sFlt-1/PlGF ratio in the assessment of preeclampsia. *Am J Obstet Gynecol* 2010;**202**(2):161 e1–e11.

144. Verlohren S, Herraiz I, Lapaire O, *et al.* The sFlt-1/PlGF ratio in different types of hypertensive pregnancy disorders and its prognostic potential in preeclamptic patients. *Am J Obstet Gynecol* 2012;**206**(1):58.e1–8.

145. Vinter CA, Jensen DM, Ovesen P, Beck-Nielsen H, Jorgensen JS. The LiP (Lifestyle in Pregnancy) study: a randomized controlled trial of lifestyle intervention in 360 obese pregnant women. *Diabetes Care* 2011;**34**(12):2502–7.

146. Vollebregt KC, Gisolf J, Guelen I, Boer K, van Montfrans G, Wolf H. Limited accuracy of the hyperbaric index, ambulatory blood pressure and sphygmomanometry measurements in predicting gestational hypertension and preeclampsia. *J Hypertens* 2010;**28**(1):127–34.

147. Widmer M, Cuesta C, Khan KS, *et al.* Accuracy of angiogenic biomarkers at 20 weeks' gestation in predicting the risk of pre-eclampsia: a WHO multicentre study. *Pregnancy Hypertens* 2015;**5**(4):330–8.

148. Wright E, Audette MC, Ye XY, *et al.* Maternal vascular malperfusion and adverse perinatal outcomes in low-risk nulliparous women. *Obstet Gynecol* 2017;**130**(5):1112–20.

149. Wright J, Small N, Raynor P, *et al.* Cohort Profile: the Born in Bradford multi-ethnic family cohort study. *Int J Epidemiol* 2013;**42**(4):978–91.

150. Zhang J, Troendle JF, Levine RJ. Risks of hypertensive disorders in the second pregnancy. *Paediatr Perinat Epidemiol* 2001;**15**(3):226–31.

151. Coomarasamy A, Williams H, Truchanowicz E, *et al.* A randomized trial of progesterone in women with recurrent miscarriages. *N Engl J Med* 2015;**373**(22):2141–8.

152. Coomarasamy A, Devall AJ, Cheed V, *et al.* A randomized trial of progesterone in women with bleeding in early pregnancy. *N Engl J Med* 2019;**380**(19):1815–24.

153. Souza RT, Cecatti JG, Costa ML, *et al.* Planning, implementing, and running a multicentre preterm birth study with Biobank resources in Brazil: the Preterm SAMBA Study. *Biomed Res Int* 2019;**2019**:5476350.

154. Hawkins TL, Roberts JM, Mangos GJ, Davis GK, Roberts LM, Brown MA. Plasma uric acid remains a marker of poor outcome in hypertensive pregnancy: a retrospective cohort study. *BJOG* 2012;**119**(4):484–92.

155. Pilalis A, Souka AP, Antsaklis P, *et al.* Screening for pre-eclampsia and small for gestational age fetuses at the 11–14 weeks scan by uterine artery Dopplers. *Acta Obstet Gynecol Scand* 2007;**86**(5):530–4.

156. Dhillon-Smith RK, Middleton LJ, Sunner KK, *et al.* Levothyroxine in women with thyroid peroxidase antibodies before conception. *N Engl J Med* 2019;**380**(14):1316–25.

157. Gabbay-Benziv R, Aviram A, Bardin R, *et al.* Prediction of small for gestational age: accuracy of different sonographic fetal weight estimation formulas. *Fetal Diagn Ther* 2016;**40**(3):205–13.

158. Anggraini D, Abdollahian M, Marion K. Foetal weight prediction models at a given gestational age in the absence of ultrasound facilities: application in Indonesia. *BMC Pregnancy Childbirth* 2018;**18**(1):436.

159. Meertens LJE, Scheepers HC, De Vries RG, *et al.* External validation study of first trimester obstetric prediction models (expect study I): research protocol and population characteristics. *JMIR Res Protoc* 2017;**6**(10):e203.

160. Crovetto F, Triunfo S, Crispi F, *et al.* First-trimester screening with specific algorithms for early- and late-onset fetal growth restriction. *Ultrasound Obstet Gynecol* 2016;**48**(3):340–8.

161. H Al Wattar B, Dodds J, Placzek A, *et al.* Mediterranean-style diet in pregnant women with metabolic risk factors (ESTEEM): a pragmatic multicentre randomised trial. *PLOS Med* 2019;**16**(7):e1002857-e.

162. Souza JP, Gulmezoglu A, Lumbiganon P, *et al.* Caesarean section without medical indications is associated with an increased risk of adverse short-term maternal outcomes: the 2004–2008 WHO Global Survey on Maternal and Perinatal Health. *BMC Med* 2010;**8**:71.

163. Souza JP, Gulmezoglu AM, Carroli G, Lumbiganon P, Qureshi Z, Group WR. The world health organization multicountry survey on maternal and newborn health: study protocol. *BMC Health Serv Res* 2011;**11**:286.

164. Zhang J, Troendle J, Reddy UM, *et al.* Contemporary cesarean delivery practice in the United States. *Am J Obstet Gynecol* 2010;**203**(4):326 e1–e10.

165. Weiner CP, Sabbagha RE, Vaisrub N, Socol ML. Ultrasonic fetal weight prediction: role of head circumference and femur length. *Obstet Gynecol* 1985;**65**(6):812–7.

166. Liu CM, Chang SD, Cheng PJ. Prediction of fetal birthweight in Taiwanese women with pre-eclampsia and gestational hypertension using an equation based on maternal characteristics. *J Obstet Gynaecol Res* 2008;**34**(4):480–6.

167. Papastefanou I, Souka AP, Eleftheriades M, Pilalis A, Chrelias C, Kassanos D. Predicting fetal growth deviation in parous women: combining the birth weight of the previous pregnancy and third trimester ultrasound scan. *J Perinat Med* 2015;**43**(4):485–92.

168. Sharp A, Jackson R, Cornforth C, *et al.* A prediction model for short-term neonatal outcomes in severe early-onset fetal growth restriction. *Eur J Obstet Gynecol Reprod Biol* 2019;**241**:109–18.

169. National Collaborating Centre for Women's and Children's Health (UK). *Antenatal Care: Routine Care for the Healthy Pregnant Woman*. London: RCOG Press; 2008 Mar. PMID: 21370514. URL: www.ncbi.nlm.nih.gov/books/NBK51886/ (accessed 12 July 2021).

170. National Institute for Health and Care Excellence (NICE). *Guide to the Processes of Technology Appraisal*. URL: www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/technology-appraisal-processes-guide-apr-2018.pdf (accessed 19 November 2021).

171. NHS Digital, '*NHS Maternity Statistics, England 2018-19 [PAS] – NHS Digital*'. NHS Maternity Statistics. 2019. URL: https://digital.nhs.uk/data-and-information/publications/statistical/nhs-maternity-statistics/2018-19 (accessed 13 May 2021).

172. Department of Health, *Reference Costs 2017/18*, Dep. Heal., 2017.

173. Netten A, Dennett J, Knight J. *Unit Costs of Health and Social Care 1998*. 1998. URL: www.rics.org/uk/knowledge/bcis/about-bcis/rebuilding/bcis-house-rebuilding-cost-index/ (accessed 12 May 2021).

174. Vieira MC, Relph S, Copas A, *et al.* The DESiGN trial (DEtection of Small for Gestational age Neonate), evaluating the effect of the Growth Assessment Protocol (GAP): study protocol for a randomised controlled trial. *Trials* 2019;**20**(1):154.

175. Pay AS, Wiik J, Backe B, Jacobsson B, Strandell A, Klovning A. Symphysis-fundus height measurement to predict small-for-gestational-age status at birth: a systematic review. *BMC Pregnancy Childbirth* 2015;**15**:22.

176. Haragan AF, Hulsey TC, Hawk AF, Newman RB, Chang EY. Diagnostic accuracy of fundal height and handheld ultrasound-measured abdominal circumference to screen for fetal growth abnormalities. *Am J Obstet Gynecol* 2015;**212**(6):820.e1–8.

177. Unit Costs of Health and Social Care 2006. PSSRU. URL: www.pssru.ac.uk/project-pages/unit-costs/unit-costs-2006/ (accessed 13 May 2021).

178. Resource impact of NICE guidance | Into practice | What we do | About | NICE. URL: www.nice.org.uk/about/what-we-do/into-practice/resource-impact-assessment (accessed 24 May 2021).

179. 3.5.7. What Are the Steps in Implementing an Impact Assessment? Marketlinks. URL: www.marketlinks.org/good-practice-center/value-chain-wiki/what-are-steps-implementing-impact-assessment (accessed 24 May 2021).

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

181. McCowan LM, Thompson JM, Taylor RS, *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**(1):e0169311.

182. Wilson ECF, Wastlund D, Moraitis AA, Smith GCS. Late pregnancy ultrasound to screen for and manage potential birth complications in nulliparous women: a cost-effectiveness and value of information analysis. *Value Health* 2021;**24**(4):513–21.

183. Smith GC, Moraitis AA, Wastlund D, *et al.* Universal late pregnancy ultrasound screening to predict adverse outcomes in nulliparous women: a systematic review and cost-effectiveness analysis. *Health Technol Assess* 2021;**25**(15):1–190.

184. Khalil A, Gordijn SJ, Beune IM, *et al.* Essential variables for reporting research studies on fetal growth restriction: a Delphi consensus. *Ultrasound Obstet Gynecol* 2019;**53**(5):609–14.

185. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol* 2015;**68**(1):25–34.

186. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J* 2021;**14**(1):49–58.

187. Riley RD, Tierney JF, Stewart LA. *Individual Participant Data Meta-analysis: A Handbook for Healthcare Research*. Hoboken: Wiley; 2021.

188. Husereau D, Drummond M, Augustovski F, *et al. Consolidated Health Economic Evaluation Reporting Standards* 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. *BMJ* 2022;**376**:e067975.

189. Philips Z, Ginnelly L, Sculpher M, *et al.* Review of guidelines for good practice in decision-analytic modelling in health technology assessment. *Health Technol Assess* 2004;**8**(36):iii–iv, ix–xi, 1–158.

190. Velauthar L, Plana MN, Kalidindi M, *et al.* First-trimester uterine artery Doppler and adverse pregnancy outcome: a meta-analysis involving 55,974 women. *Ultrasound Obstet Gynecol* 2014;**43**(5):500–7.

191. Sibai BM. Management of late preterm and early-term pregnancies complicated by mild gestational hypertension/pre-eclampsia. *Semin Perinatol* 2011;**35**(5):292–6.

192. Shah A, Faundes A, Machoki M, *et al.* Methodological considerations in implementing the WHO Global Survey for Monitoring Maternal and Perinatal Health. *Bull World Health Organ* 2008;**86**(2):126–31.

193. Souza JP; WHO Multicountry Survey on Maternal and Newborn Health Research Network. The World Health Organization Multicountry Survey on Maternal and Newborn Health project at a glance: the power of collaboration. *BJOG* 2014;**121** (Suppl. 1):v–viii.

194. Eunice Kennedy Shriver National Institute of Child Health and Human Development. *NICHD DASH Consortium of Safe Labor (CSL) Study Page*. URL: https://dash.nichd.nih.gov/study/2331 (accessed 25 July 2018).

195. Souza RT, Costa ML, Mayrink J, *et al.* Clinical and epidemiological factors associated with spontaneous preterm birth: a multicentre cohort of low risk nulliparous women. *Sci Rep* 2020;**10**(1):855.

Appendix 1 Detailed study characteristics of IPPIC cohorts

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
IPPIC UK							
SCOPE[119]	Observational	Prospective cohort	Multicountry (UK, New Zealand, Australia and Republic of Ireland)	2004–8	Low risk	Healthy nulliparous women with singleton pregnancies	Recognised as high risk of PE, SGA baby or spontaneous preterm birth due to underlying medical condition such as chronic hypertension requiring antihypertensive drugs, diabetes, renal disease, systemic lupus erythematosus, antiphospholipid syndrome, sickle cell disease or HIV. Previous cervical knife cone biopsy, three or more abortions or miscarriages, current ruptured membranes, known major fetal anomaly or abnormal karyotype and interventions that can alter the course of pregnancy such as aspirin or cervical suture
Allen[80]	Observational	Prospective cohort	UK	2010–4	Any pregnancy	All pregnant women attending an inner London hospital between 11 and 14 weeks' gestation	Women with multiple pregnancies and fetal anomalies
ALSPAC[75]	Observational	Prospective birth cohort	UK	1991–2	Any pregnancy	All pregnant women resident in Avon, UK	None
Chappell[93]	Randomised	Trial	UK	NI	High risk	Pregnant women with an abnormal Doppler waveform in either uterine artery at 18–22 weeks' gestation or a history of preeclampsia in a previous pregnancy which led to preterm delivery, eclampsia or HELLP syndrome	Heparin or warfarin treatment, abnormal fetal-anomaly scan or multiple pregnancy

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
EMPOWAR[94]	Randomised	Trial	UK	2011–4	High risk	Women at least 16 years of age at recruitment, between 12 and 16 weeks' gestation and with a BMI of 30 kg/m²	Non-white women and those with: history of diabetes, systemic disease at the time of enrolment (requiring either regular drugs or systemic corticosteroids treatment in the past 3 months), previous delivery of a baby smaller than the 3rd centile for weight, history of PE with delivery before 32 weeks' gestation, known hypersensitivity to metformin hydrochloride or any of the excipients. Known liver or renal failure, acute disorders at the time of trial entry with the potential to change renal function, such as dehydration sufficient to require intravenous infusion, severe infection, shock, intravascular administration of iodinated contrast agents or acute or chronic diseases that might cause tissue hypoxia (e.g. cardiac or respiratory failure, recent myocardial infarction, hepatic insufficiency, acute alcohol intoxication or alcoholism); lactating women; and women with multiple pregnancy
POPPY[78]	Observational	Prospective cohort	UK	2011–3	Any pregnancy	Pregnant asymptomatic women with a high risk of spontaneous preterm birth, such as previous history of spontaneous preterm delivery, late miscarriage, invasive cervical surgery or a short cervix	NI

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Poston 2006[123]	Randomised	Trial	UK	2003–5	High risk	Gestational age 14–21 weeks plus one or more of the following risk factors: history of preeclampsia in preceding requiring preterm delivery, history of HELLP syndrome, eclampsia, essential hypertension requiring medication, maternal diastolic blood pressure of 90 mm Hg or more before 20 weeks' gestation in the current pregnancy, history of diabetes, antiphospholipid syndrome; 8 chronic renal disease, multiple pregnancy; abnormal uterine artery doppler waveform, primiparity with (BMI at first antenatal appointment of 30 kg/m^2 or more	Women taking vitamin supplements containing doses of vitamin C of 200 mg or more or of vitamin E of 40 IU or more daily. Women treated with warfarin
Poston 2015[122]	Randomised	Trial	UK	2009–14	High risk	Women older than 16 years with a BMI of 30 kg/m^2 or higher and a singleton pregnancy	Any underlying disorders, including a pre-pregnancy diagnosis of essential hypertension, diabetes, renal disease, systemic lupus erythematosus, antiphospholipid syndrome, sickle cell disease, thalassaemia, coeliac disease, thyroid disease and current psychosis; or if on metformin
Macleod[109]	Randomised	Trial	UK	NI	High risk	Women identified to be at high risk of adverse pregnancy outcome by uterine arterial waveform analysis	Women with underlying conditions thought likely to compromise renal function such as diabetes or renal disease
St George[135]	Observational	Prospective registry	UK	2000–15	Any pregnancy	All pregnant women attending an inner London hospital	None
PARIS[84]	IPD MA of 31 randomised trials	Trial	33 countries	1985–2005	Varied	Varied (dependent on individual study)	Varied (dependent on individual study)

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
AMND[86]	Observational	Prospective registry	UK	1986–2015	Any pregnancy	Data from every pregnancy event occurring in Aberdeen Maternity Hospital	None
BIB[149]	Observational	Prospective birth cohort	UK	2007–11	Any pregnancy	All pregnant women attending Bradford Royal Infirmary	None
PROMISE[151]	Randomised	Trial	UK	2010–3	Any pregnancy	Women with history of unexplained miscarriage who conceived within study period	Any thrombophilic condition, uterine cavity abnormalities, diabetes, thyroid disease, SLE, on heparin treatment or contraindicated to progesterone
PRISM[152]	Randomised	Trial	UK	2015–7	Any pregnancy	Women < 12 weeks pregnant with vaginal bleeding no older than 39 years	CRL ≥ 7mm with no heartbeat, ectopic pregnancy, life-threatening bleeding and contraindication to progesterone use
Velauthar[190]	Observational	Prospective cohort	UK	NI	Any pregnancy	All pregnant women attending an inner London hospital	None
TABLET[156]	Randomised	Trial	UK	2011–6	Any pregnancy	Pregnant women 16–40 years with previous miscarriage or on treatment for infertility	Women receiving treatment for thyroid disease, had cardiac disease or were on lithium or amiodarone
ESTEEM[161]	Observational	Prospective cohort	UK	2014–6	Any pregnancy	Singleton pregnancies <18 weeks gestation with proficient English language ability	Pre-existing diabetes, gestational diabetes, chronic renal disease, autoimmune disease, on statins or similar drugs
POP[18]	Observational	Prospective cohort	UK	2008–12	Any pregnancy	Nulliparous women with singleton pregnancies	None

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
IPPIC International							
Baschat[87]	Observational	Prospective cohort	USA	2007–10	Any pregnancy	All pregnant women attending any of 4 Baltimore (USA) hospitals for first trimester screening	None
Audibert[85]	Observational	Prospective cohort	Canada	2006–8	Low risk	Nulliparous women with singleton pregnancies presenting for Down syndrome screening at 11–13 weeks	Pregnancies with a major fetal chromosomal or structural anomaly
Caradeux[90]	Observational	Prospective cohort	Chile	NI	Any pregnancy	All pregnant women attending for an 11–14 week ultrasound evaluation	None
Giguere[98]	Observational	Prospective cohort	Canada	2005–10	Any pregnancy	Women at least 18 years old and with a gestational age of at least 10 weeks at their first prenatal visit with no chronic hepatic or renal diseases	Pregnancies with major fetal abnormalities and those ending in termination, miscarriage or fetal death before 24 weeks of gestation
Goetzinger[100]	Observational	Retrospective cohort	USA	2003–8	Any pregnancy	Women seen for aneuploidy screening	None
Antsaklis[82]	Observational	Prospective cohort	Greece	1997–8	Low risk	All nulliparous women	Women with multiple pregnancies, renal disease, cardiovascular diseases and fetal anomalies
Llurba[112]	Observational	Prospective cohort	Spain	2002–6	Any pregnancy	Singleton women attending routine second trimester anomaly scans	None
WHO[147]	Observational	Prospective cohort	Multicountry (Argentina, Colombia, India, Italy, Kenya, Peru, Switzerland and Thailand)	2006–9	High risk	Women with risk factors for PE	Women with known renal disease or proteinuria

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Andersen[81]	Observational	Prospective cohort	Denmark	2010–2	Any pregnancy	Newly pregnant women	Twin pregnancies and early pregnancy fetal losses
Arenas[83]	Observational	Prospective cohort	Spain	2000–1	Any pregnancy	Women attending routine ultrasound scan at 20 weeks	Multiple pregnancies or congenital defects
FINNPEC[105]	Observational	Prospective/ retrospective case-control cohort	Finland	2008–11	Any pregnancy	Nulliparous or multiparous women with a singleton pregnancy with or without PE on admission to hospital	Multiple pregnancy, maternal age < 18 years
Galindo[143]	Observational	Prospective case-control cohort	Spain	NI	Any pregnancy	Singleton pregnancies	Multigestation, antiphospholipid antibody syndrome, systemic lupus erythematosus or any other auto-immune disease as well as chronic corticosteroid or non-steroidal anti-inflammatory drug use except low-dosage aspirin < 150 mg/day
Generation R[106]	Observational	Prospective birth cohort	The Netherlands	2002–6	Any pregnancy	Resident mothers delivering in the study period	None
NICHD HR[191]	Randomised	Trial	USA	1991–5	High risk	Women with pregestational, insulin-treated diabetes mellitus, women with chronic hyperten-sion, women with multifetal gestations and women who had had preeclampsia in a previous pregnancy	Women with multifetal gestation if they also had chronic hypertension, renal disease, diabetes, history of PE and current proteinuria
NICHD LR[132]	Randomised	Trial	USA	NI (early 1990s)	Low risk	Healthy nulliparous women	Women with chronic hypertension, renal disease, diabetes and other illnesses

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Placental Health Study[148]	Observational	Prospective cohort	Canada	2012–3	Low risk	Healthy nulliparous women with singleton pregnancies	Chronic hypertension, use of unfractionated or low-molecular-weight heparin, pregestational diabetes mellitus, major fetal abnormalities, ruptured membranes, vaginal bleeding from 13 0/7 weeks of gestation for >1 day or a short cervical length on ultrasonography before 20 weeks of gestation (<2 cm long)
POUCH[103]	Observational	Prospective cohort	USA	1998–2004	Any pregnancy	Women with a singleton pregnancy at 16–27 weeks' gestation, no known chromosomal abnormality, maternal age of at least 15 years, no pre-pregnancy diabetes mellitus	None
Van kuijk 2011[139]	Observational	Prospective cohort	The Netherlands	1993–2008	High risk	Women with preceding singleton pregnancy complicated by PE or HELLP syndrome	NI
Van kuijk 2014[138]	Observational	Prospective and retrospective cohort	The Netherlands	2008–12	High risk	Women with preceding singleton pregnancy complicated by PE or HELLP syndrome	Women who had diabetes, autoimmune disease, heart or kidney disease
Odibo[120]	Observational	Prospective cohort	USA	2009–11	Any pregnancy	Women attending first trimester screening	None
PREDO[99]	Observational	Prospective case-control cohort	Finland	2005–9	Any pregnancy	Pregnant women with known risk factor for preeclampsia and IUGR and those without, attending clinics for their first ultrasound screening between 12 and 14 weeks gestation	Asthma diagnosed by a physician, allergy to ASA, tobacco smoking during pregnancy, previous peptic ulcer, previous placental ablation, inflammatory bowel diseases (Crohn's disease, ulcerative colitis), rheumatoid arthritis, haemophilia or thrombophilia (previous venous or pulmonary thrombosis and/or coagulation abnormality), gestational weeks + days <12 + 0 or more than 14 + 0 or multiple pregnancy

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Prefumo[124]	Observational	Prospective cohort	Italy	2001–5	Any pregnancy	Women attending routine antenatal care	Known medical condition (e.g. diabetes mellitus, connective tissue disease, essential hypertension) or a history of recurrent miscarriage
Skrastad[133]	Observational	Prospective cohort	Norway	2010–2	High risk	Nulliparous and high-risk parous women with one or more previous PE pregnancies	Use of any anticoagulant medication or acetylsalicylic acid in pregnancy
Verlohren[144]	Observational	Prospective case-control cohort	Germany	NI	Any pregnancy	Singleton pregnancies	Multigestation, antiphospholipid antibody syndrome, systemic lupus erythematosus or any other autoimmune disease as well as chronic corticosteroid or non-steroidal anti-inflammatory drug use except low-dosage aspirin < 150 mg/day
Rumbold[128]	Randomised	Trial	Australia	2001–5	Low risk	Nulliparous women with a singleton pregnancy between 14 and 22 weeks of gestation and normal blood pressure	Known multiple pregnancy, known potentially lethal fetal anomaly, known thrombophilia, chronic renal failure, antihypertensive therapy or specific contraindications to vitamin C or E therapy such as haemochromatosis or anticoagulant therapy
Vollebregt[146]	Observational	Prospective cohort	The Netherlands	2004–6	High risk	Healthy nulliparous women at low risk and women with elevated risk for preeclampsia or FGR with singleton pregnancies	NI
JSOG[77]	Observational	Prospective registry	Japan	2013–4	Any pregnancy	All women giving birth at participating institutions in Japan	None

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
DOMINO[115]	Randomised	Trial	Australia	2005–8	Any pregnancy	Singleton pregnancies at <21 weeks' gestation	Already taking a prenatal supplement with DHA, their fetus had a known major abnormality, they had a bleeding disorder in which tuna oil was contraindicated, were taking anticoagulant therapy, had a documented history of drug or alcohol abuse, were anticipating in another fatty acid trial
Danish Birth Cohort[108]	Observational	Prospective registry	Denmark	1996–2002	Any pregnancy	All women in Denmark	None
Indonesian cohort[130]	Observational	Prospective cohort	Indonesia	2012–5	Any pregnancy	All women attending antenatal care	None
Ohkuchi[121]	Observational	Prospective cohort	Japan	NI	Any pregnancy	Women with singleton pregnancies attending antenatal care	None
Lecarpentier[111]	Observational	Retrospective cohort	France	2004–7	High risk	Women with chronic hypertension	Multiple pregnancies, women with secondary hypertension, women with proteinuria at <20 weeks' gestation, women considered as having a chronic hypertension but without any treatment at first prenatal visit, women transferred from other maternities, pregnancies complicated by fetal malformations
TEST[118]	Randomised	Trial	Ireland	2014–6	Low risk	Nulliparous women between 11 and 14 weeks gestation and not already on aspirin	Fetal abnormality or contraindication to aspirin
Masse[116]	Observational	Prospective cohort	Canada	1989–91	Low risk	Nulliparous women attending hospital for routine blood sampling at the start of pregnancy	Diabetes mellitus, cardiovascular disease (including chronic hypertension) or renal disease or women seen after 20 weeks

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Staff[134]	Observational	Prospective case-control cohort	Norway	NI	Low risk	Women with singleton pregnancies	NI
STORK G[107]	Observational	Prospective cohort	Norway	2008–10	Any pregnancy	Healthy pregnant women	Women with diabetes or diseases require intensive hospital follow-up in pregnancy
Vatten[142]	Observational	Prospective case-control cohort	Norway	1992–4	Any pregnancy	Women attending antenatal care	None
Vinter[145]	Randomised	Trial	Denmark	2007–10	High risk	Women aged between 18 and 40 with a pre-pregnancy weight of between 30 and 45 kg/m^2	Women with chronic medical disorders (hypertension, diabetes, alcohol or drug use) and serious obstetric complication (multiple pregnancy, congenital malformation, miscarriage)
BORN Ontorio[76]	Observational	Prospective registry	Canada	2012–4	Any pregnancy	Women giving birth during the data period in the Ontario region	None
Ghana Cohort[137]	Observational	Prospective cohort	Ghana	2012–4	Any pregnancy	Women < 17 weeks pregnant, at least 18 years old with no established hypertension at booking	None
MoBA[114]	Observational	Prospective registry	Norway	1999–2005	Any pregnancy	All women giving birth in Norway	None
Huang[104]	Observational	Retrospective cohort	Canada	2000–3	Any pregnancy	All women screened in early pregnancy for Down syndrome	None
Carbillion[91]	Observational	Prospective registry	France	1996–2005	Any pregnancy	Women giving birth in the data period in that region	None

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Goffinet[101]	Randomised	Trial	France	1994–7	Low risk	All women attending routine antenatal visit before 24 weeks	Any indications for UAD such as chronic hypertension, diabetes, previous fetal death, IUGR, hypertensive disorders of pregnancy or contraindication for aspirin
Rang[125]	Observational	Prospective cohort	The Netherlands	NI	High risk	Women with a history of early-onset preeclampsia in a previous pregnancy or women who had never been pregnant	None
Cameroni 2011[89]	Observational	Retrospective cohort	Italy	NI	High risk	Singleton pregnancies at risk of PE or IUGR	NI
Conserva 2012[95]	Observational	Prospective cohort	Italy	2001–8	High risk	Women with previous adverse pregnancy outcomes	Multiple gestation; a previous uneventful pregnancy; a previous pregnancy treated with LMWH or unfractionated heparin; patients with clinical immune disease and acquired thrombophilia – lupus-like anticoagulant or APL syndrome; patients with positive antinuclear, antimitochondria, antismooth muscle antibodies; postnatal or post-mortem diagnosis of congenital fetal anomaly or fetal infection; women of non-Caucasian ethnicity; alcohol or illicit drug use; early pregnancy loss was not considered an APO
Facchinetti[96]	Observational	Prospective cohort	Italy	2001–6	High risk	Previous singleton pregnancies complicated by PE and received evaluation for thrombophilia	History of thromboembolic diseases, renal and/or cardiovascular disorder, systemic lupus erythematosus, diabetes and any ethnic group other than white

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Ferrazzani[131]	Observational	Prospective cohort	Italy	1990–2001	High risk	Previous severe preterm PE	Previous HELLP syndrome
Figueiro-Filho[97]	Observational	Prospective case-control cohort	Brazil	2007–10	High risk	Women with severe PE in previous pregnancies	Antiphospholipid antibodies and thrombophilia
Langenveld[110]	Observational	Retrospective cohort	The Netherlands	1996–2004	High risk	Women with hypertension (including patients with chronic hypertension), PE or HELLP syndrome, and delivered before 34 weeks of gestation in the study period and primiparous with singleton pregnancy without fetal abnormalities in first pregnancy	NI
Lykke[113]	Observational	Prospective registry	Denmark	1978–2007	Any pregnancy	Singleton deliveries of women with first delivery > 15 years and second delivery < 50 years	Cardiovascular diagnosis and type 1 or 2 diabetes
Mbah[117]	Observational	Prospective registry	USA	1989–2005	Any pregnancy	Women with first and second singleton pregnancies within the gestational age range of 20–44 weeks	None
Trogstad[136]	Observational	Prospective registry	Norway	1967–98	Any pregnancy	Women with a first and a second delivery	None
Salim[129]	Observational	Prospective cohort	Israel	2000–6	High risk	Previous pregnancy with antepartum complications at ≥23 weeks gestation	Women who had a previous pregnancy with antepartum complications that could be attributed to multiple gestations, having fetuses with major congenital anomalies or chromosomal abnormalities, fetal infection, chorioamnionitis, hydrops fetalis and diabetes mellitus

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Van Oostwaard 2012[140]	Observational	Prospective cohort	The Netherlands	2000–2	High risk	Women with a hypertensive disorder in the index pregnancy and delivery at 34–37 weeks of gestation	Fetal abnormalities
Van Oostwaard 2014[141]	Observational	Retrospective cohort	The Netherlands	2000–2	High risk	Women with a hypertensive disorder in the index pregnancy and delivery at 34–37 weeks of gestation	Fetal abnormalities
Zhang[150]	Observational	Prospective cohort	USA	1959–65	Any pregnancy	Women attending prenatal care	None
Brown 2007[88]	Observational	Retrospective cohort	Australia	1988–98	Any pregnancy	Women referred for management of hypertensive disorders of pregnancy	None
Costa 2014[102]	Observational	Prospective cohort	Brazil	2009–11	Any pregnancy	Women attending for first trimester Down syndrome screening	None
Costa 2016_1[126]	Observational	Prospective cohort	Brazil	2009–14	Any pregnancy	Women with singleton pregnancies attending for first trimester ultrasound scans	Prior maternal renal disease, major fetal malformations or chromosomal abnormalities, miscarriage
Costa 2017_1[79]	Observational	Prospective cohort	Australia	2012–5	Any pregnancy	Women attending for their second trimester morphology ultrasound between 19 and 22 weeks	None
Costa 2017_2[127]	Observational	Prospective cohort	Brazil	2009–14	Any pregnancy	Singleton pregnancies of women attending routine ultrasound screening	Kidney disease diagnosis in their previous history or on ultrasound examination, major fetal malformations or chromosomal abnormalities and fetuses with crown-rump length longer than 84 mm

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
WHO GS[192]	Observational	Prospective cohort	Multicountry (Afghanistan, Angola, Argentina, Brazil, Cambodia, China, Democratic Republic of the Congo, Ecuador, India, Japan, Jordan, Kenya, Lebanon, Mexico, Mongolia, Nepal, Nicaragua, Niger, Nigeria, occupied Palestinian territory, Pakistan, Paraguay, Peru, Philippines, Qatar, Sri Lanka, Thailand, Uganda, Vietnam)	2004–8	Any pregnancy	Pregnant women attending hospitals from Americas, Africa, Southeast Asia and Western Pacific WHO regions	None

continued

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
WHO MCS[193]	Observational	Prospective cohort	Multicountry (Algeria, Angola, Democratic Republic of the Congo, Kenya, Niger, Nigeria, Uganda, Argentina, Brazil, Cuba, Ecuador, Mexico, Nicaragua, Paraguay, Peru, Cambodia, China, India, Japan, Nepal, Philippines, Sri Lanka, Thailand, Vietnam)	2004–8	Any pregnancy	Pregnant women attending hospitals from Americas, Africa, Southeast Asia and Western Pacific WHO regions	None
Crovetto[160]	Observational	Prospective cohort	Spain	2007–12	Any pregnancy	Singleton pregnancies attending routine first trimester screening	Major fetal defects, miscarriage and termination of pregnancies without medical indication
NICHD CSL[194]	Observational	Retrospective cohort	USA	2002–8	Any pregnancy	All deliveries ≥ 23 weeks gestation from 19 hospitals across the USA	None
Expect[159]	Observational	Prospective cohort	The Netherlands	2013–5	Any pregnancy	Adult pregnant women < 16 weeks	Miscarriage and termination < 24 weeks
Anggraini[158]	Observational	Retrospective cohort	Indonesia	2013–5	Any pregnancy	Pregnant women who received antenatal care	NI

Study/dataset	Study design (randomised, observational)	Data source (trial, cohort, registry)	Country	Data period	Population type [any pregnancy, high risk (women with complications), low risk]	Inclusion criteria	Exclusion criteria
Gabby-Benziv[157]	Observational	Retrospective cohort	Israel	2007–14	Any pregnancy	All singleton pregnant women attending for ultrasound scan	None
Pilalis[155]	Observational	Prospective cohort	Greece	NI	Any pregnancy	Women with singleton pregnancies attending ultrasound examination at 11–14 weeks	NI
Souka	Observational	Prospective registry	Greece	NI	Any pregnancy	All pregnant women attending a private fetal medical centre	None
Souka 2	Observational	Prospective registry	Greece	NI	Any pregnancy	All pregnant women attending a private fetal medical centre	None
Hawkins[154]	Observational	Prospective registry	Australia	2000–8	High risk	Hypertensive pregnancies referred for renal consultation	Non-hypertensive pregnancies and women with type 1 diabetes
SAMBA[195]	Observational	Prospective cohort	Brazil	2015–8	Low risk	Nulliparous singleton pregnant women < 21 weeks gestation	≥3 abortions, chronic hypertension requiring treatment, diabetes or renal disease, arterial BP > 160/100, autoimmune disease, sickle cell disease, HIV, fetal malformation, cervical suture or knife cone biopsy, Mullerian anomalies, use of corticosteroids, aspirin, calcium, fish oil, vitamin C/E or heparin

APL, anti phospholipid; BMI, body mass index; DHA, docosahexaenoic acid; IUGR, intra uterine growth restriction; LMWH, low molecular weight heparin; UAD, uterine artery doppler.

Appendix 2 Prediction study Risk of bias assessment (RoB)[a] of cohorts on the IPPIC Network database used for external validation and model development

Study	Participants	Predictors	Outcome	Overall RoB
Allen	+	+	+	+
ALSPAC	+	+	+	+
Baschat	+	+	+	+
Generation R	+	+	+	+
Odibo	+	+	+	+
Rumbold	+	+	+	+
JSOG	+	?	+	?
STORKG	+	+	+	+
POP	+	+	+	+
NICHD CSL	+	+	+	+

a + indicates low RoB; – indicates high RoB; ? indicates unclear RoB.

Appendix 3 Predicted birthweight distribution

Predicted birthweight was slightly skewed, with a long left tail, in all included datasets. While all datasets recorded observed birthweights in this left tail, almost down to zero, very few babies were born with actual birthweight over 5000 g (5 kg). The largest babies were seen in JSOG, where 27 babies were born larger than 5 kg (potentially a reflection of the larger size of this dataset, allowing more extreme observations to be seen). While the left tail of the observed distribution was well modelled by Poon 2011, the more extreme right observations were poorly identified, with very few predicted birthweights exceeding 4000 g (4 kg) in any data set.

Overall distributions of predictions were similar across datasets, as was the distribution of observed birthweights. The model reasonably mimics the distribution of the observed outcome, where the majority of babies were born at a larger, healthier weight with gradually fewer small babies.

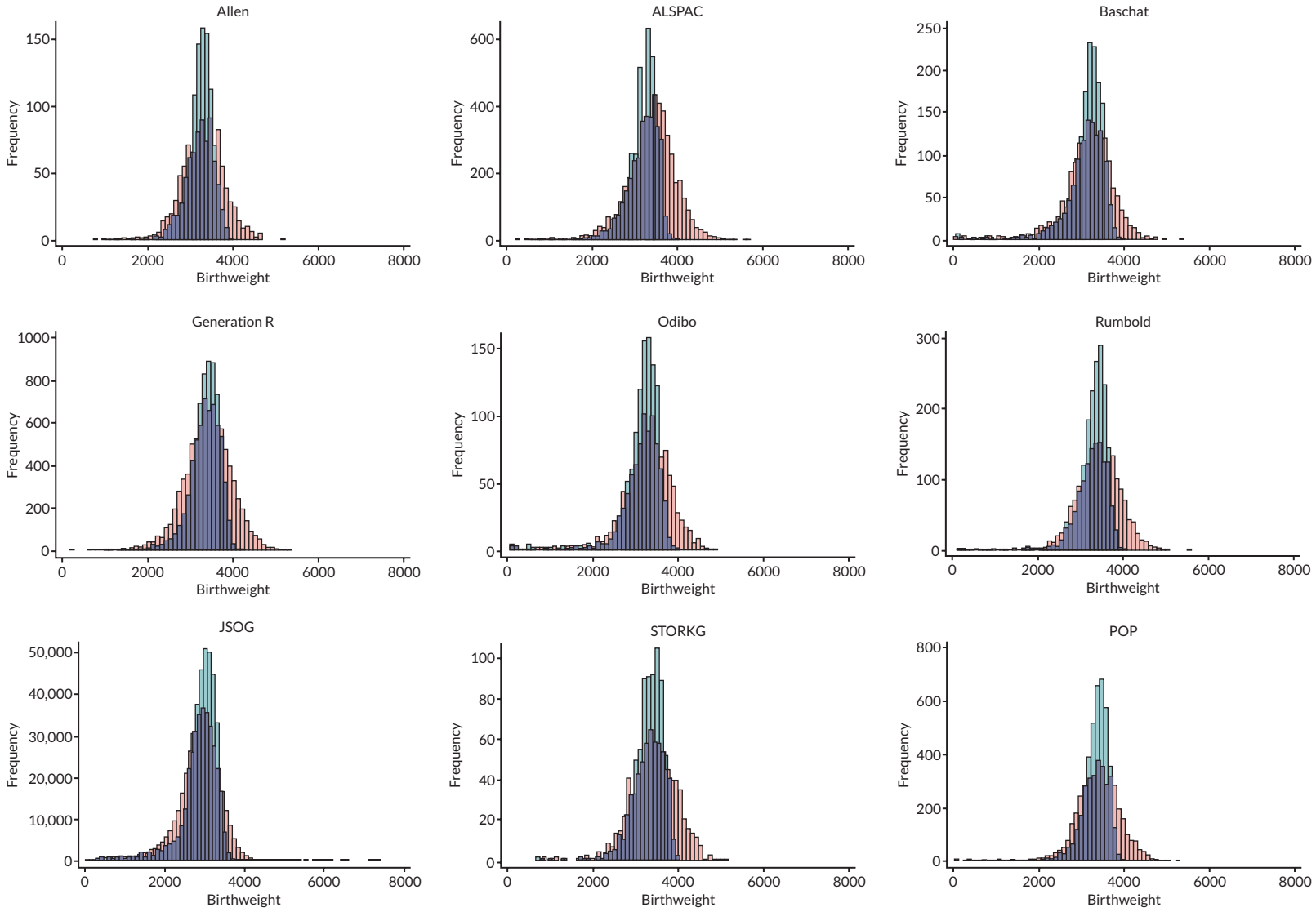

FIGURE 21 Distributions of expected (green) and observed (purple) birthweights (g), by study.

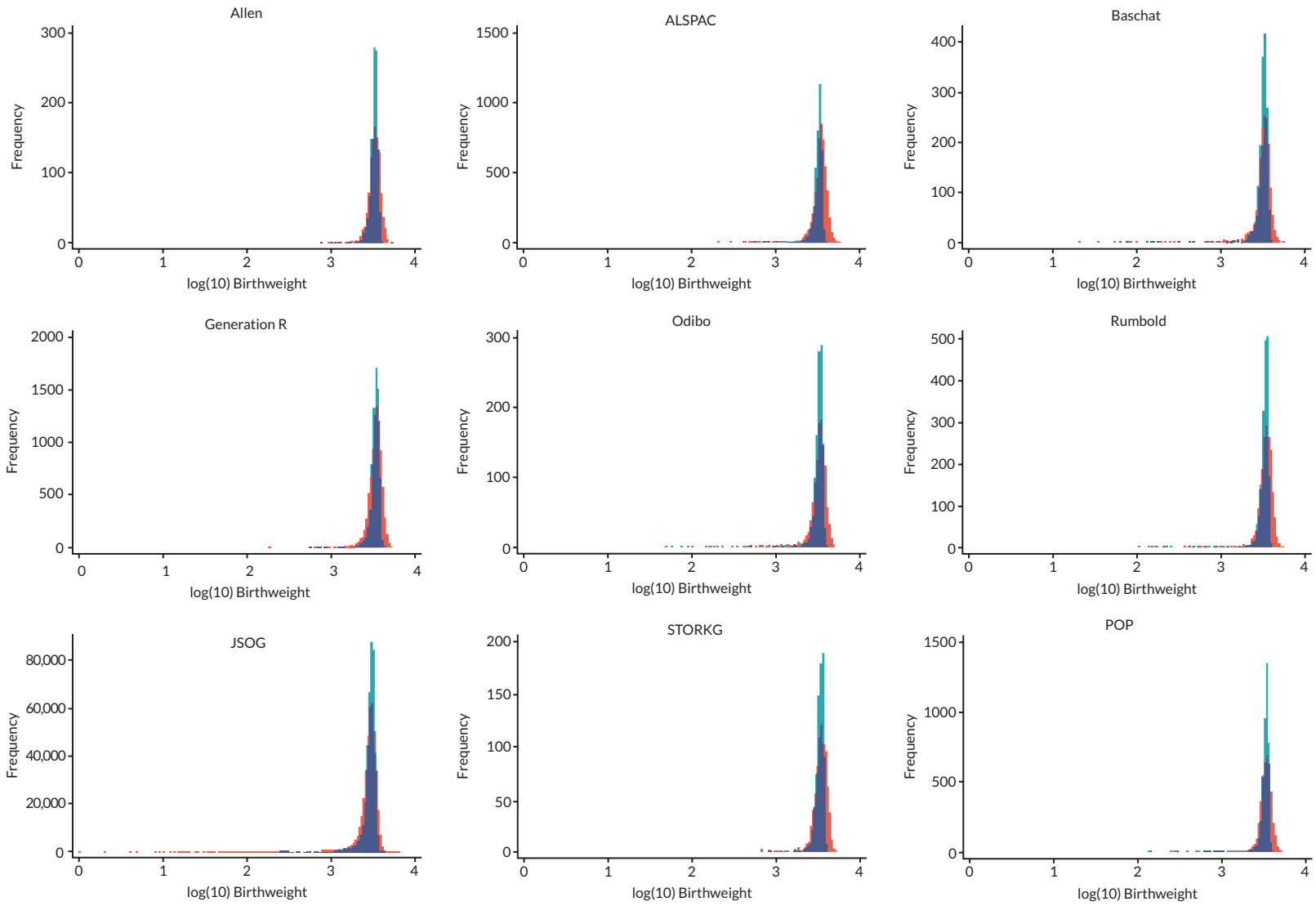

FIGURE 22 Distributions of expected (blue) and observed (red) $\log_{10}$ birthweight, by study.

Appendix 4 Summary of predictors across model development cohorts

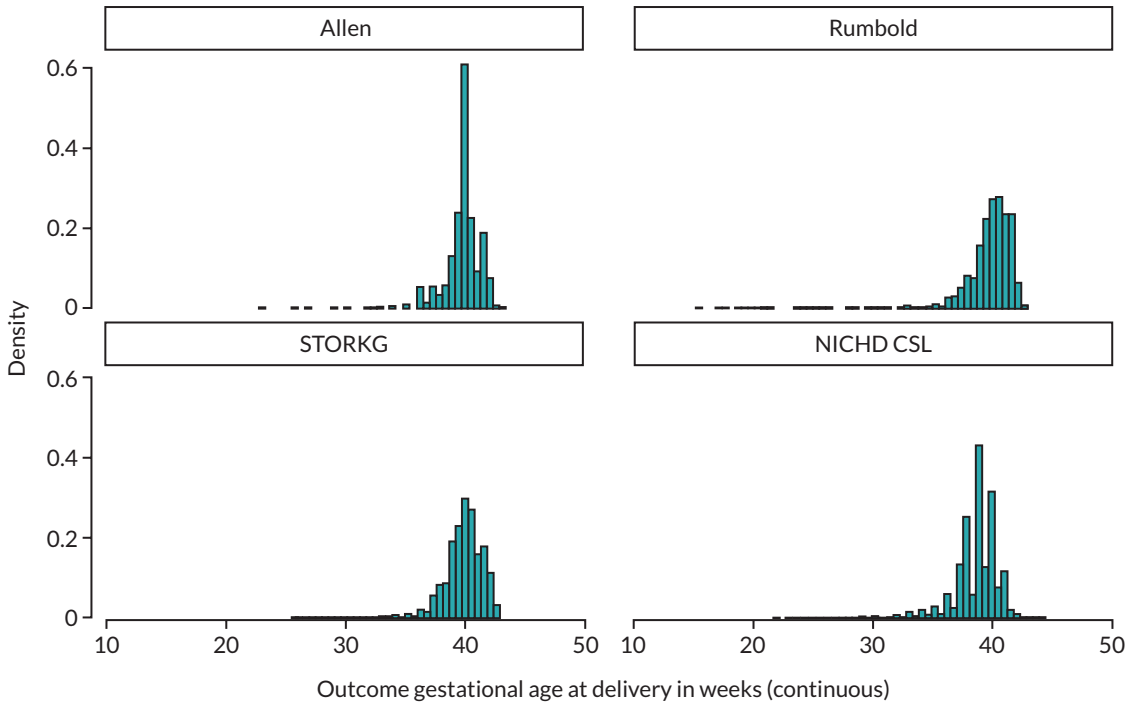

FIGURE 23 Gestational age at delivery.

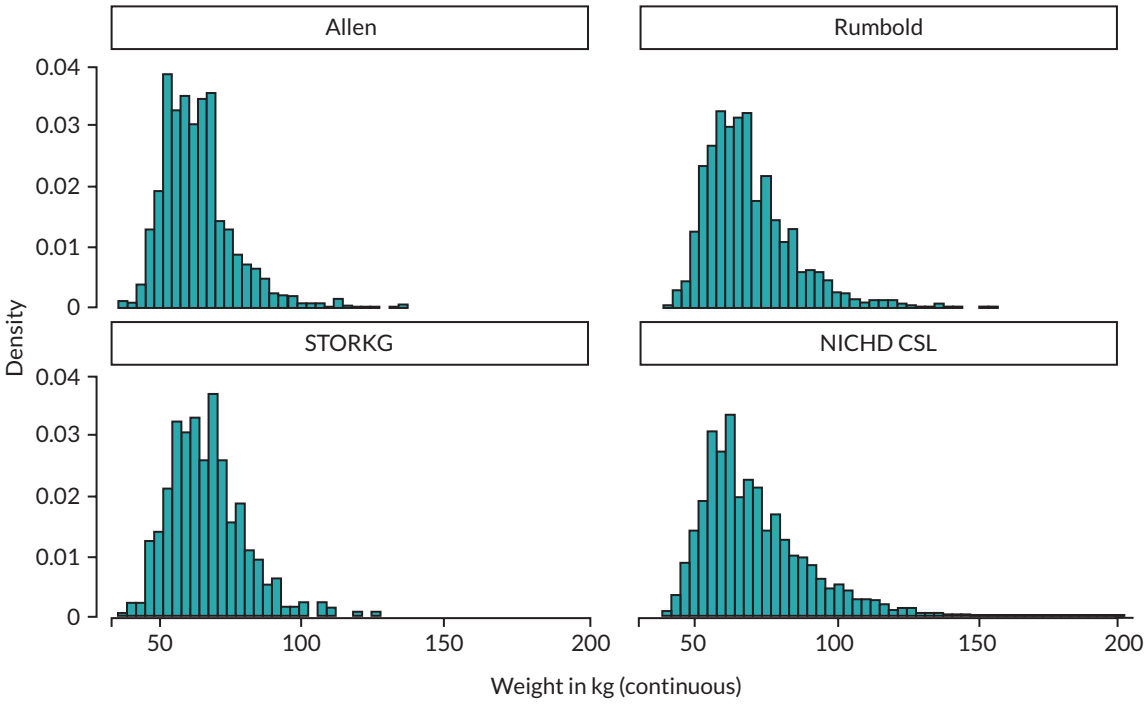

FIGURE 24 Mother's weight.

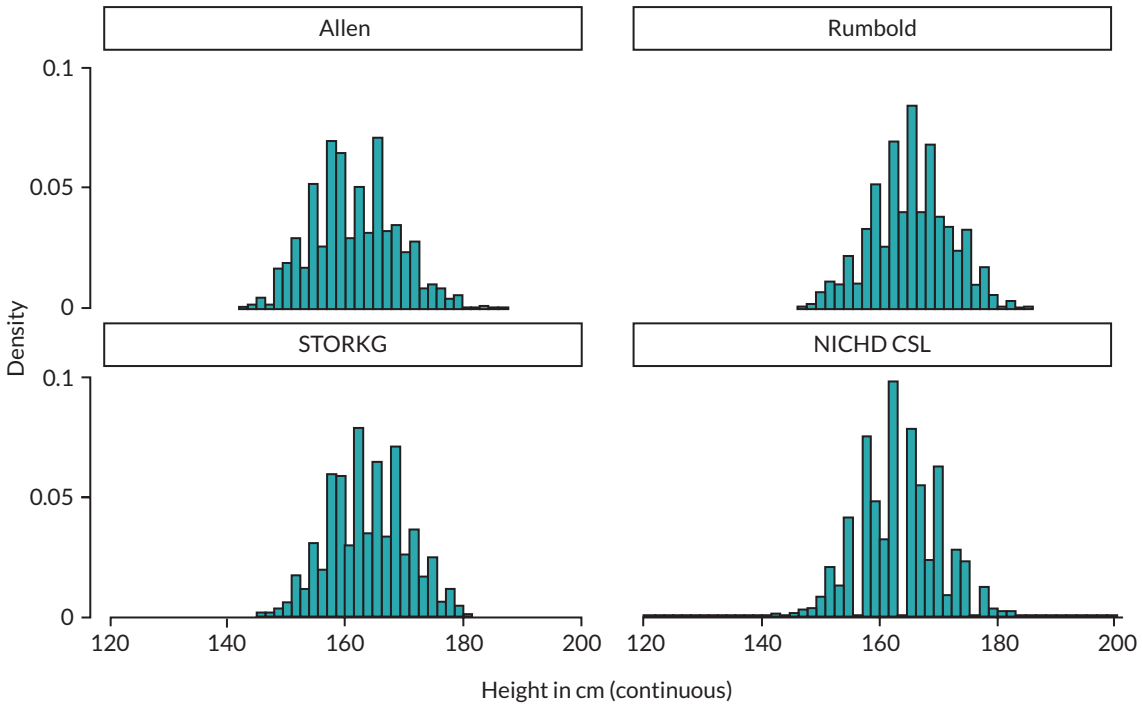

FIGURE 25 Mother's height.

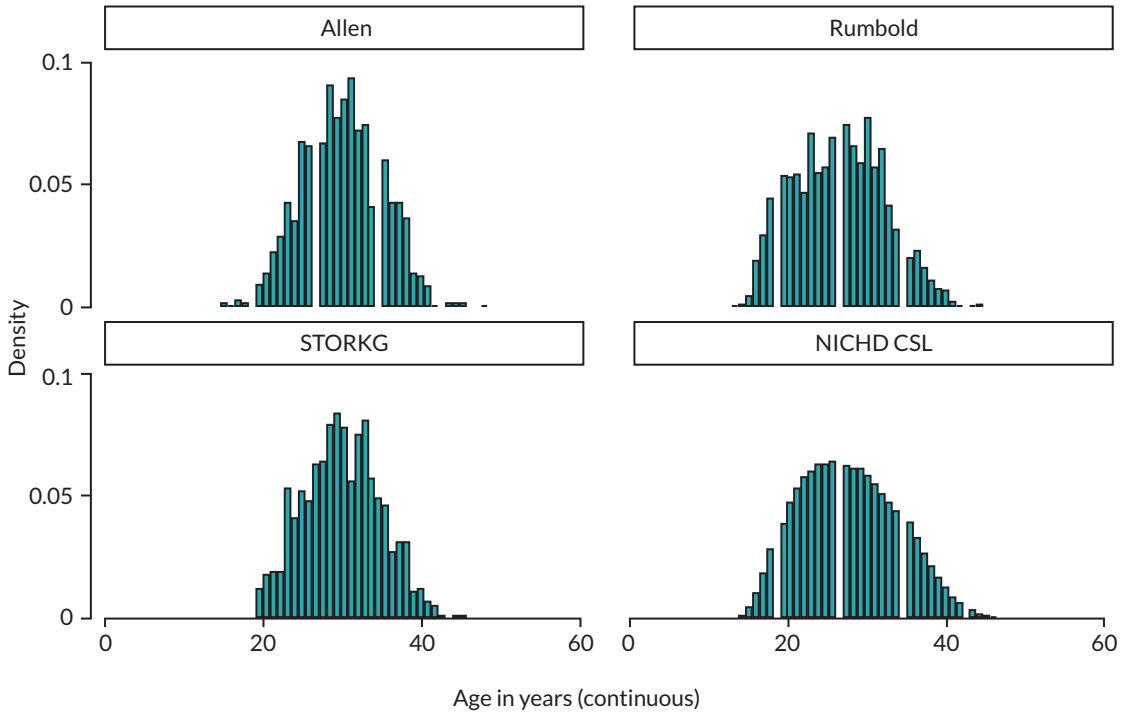

FIGURE 26 Mother's age.

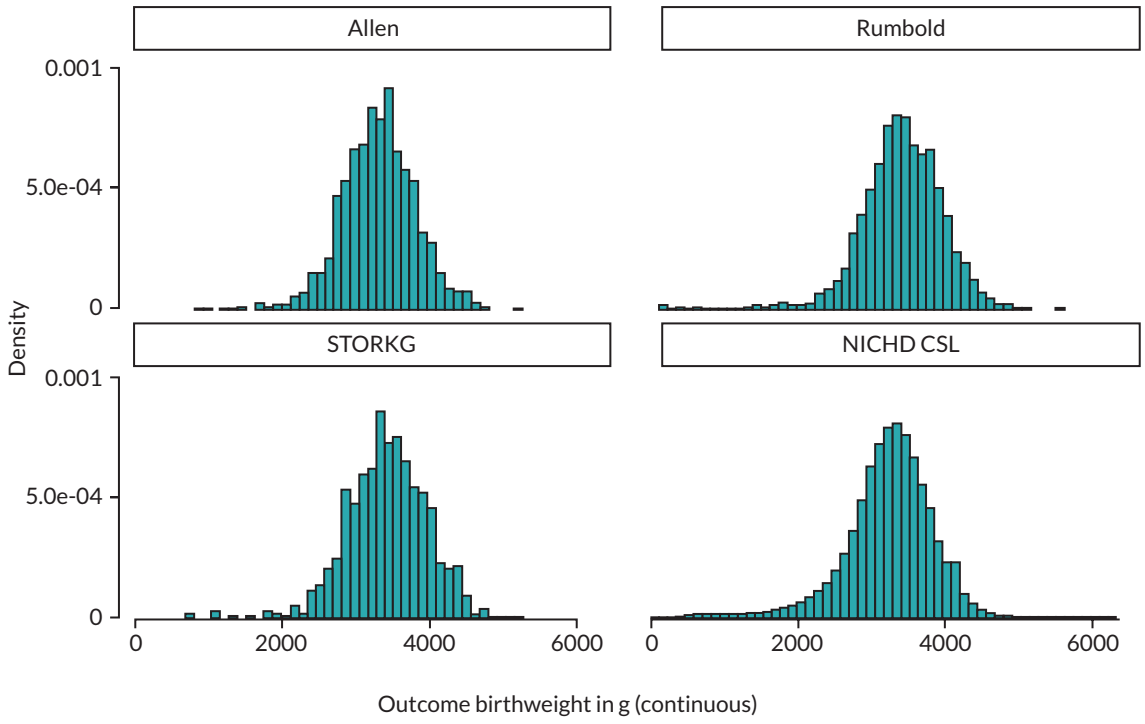

FIGURE 27 Birthweight.

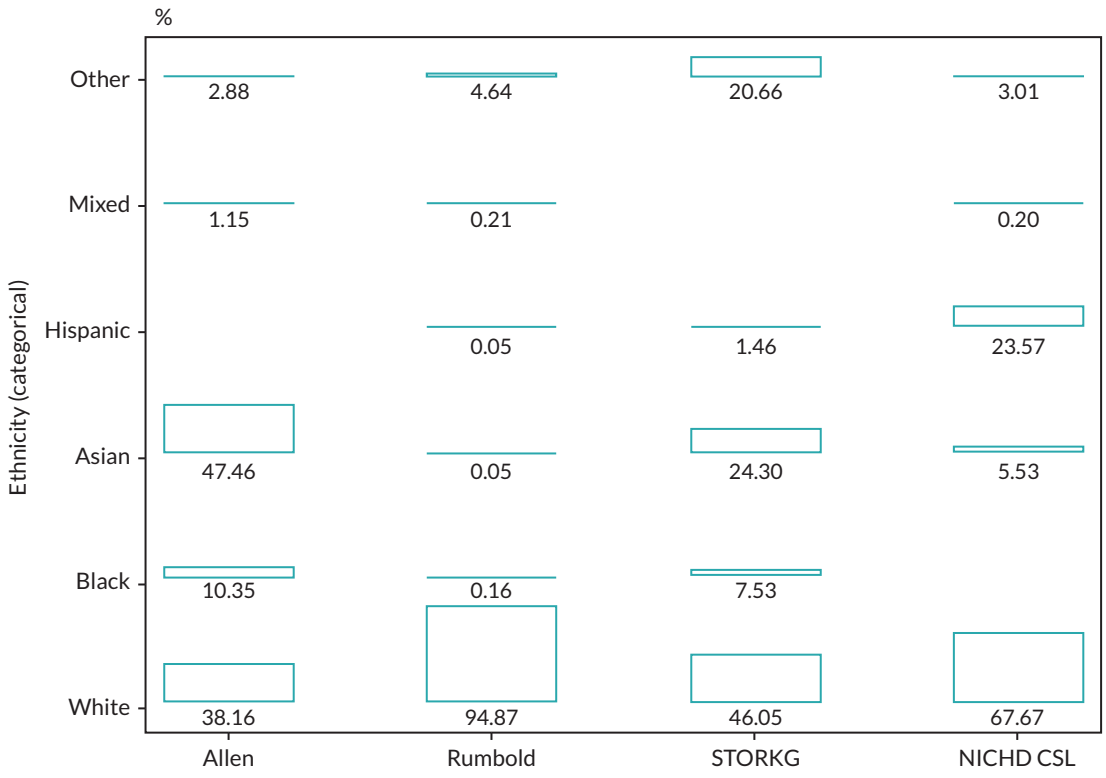

FIGURE 28 Ethnicity.

Appendix 5 Imputation checking for model development

Continuous variables

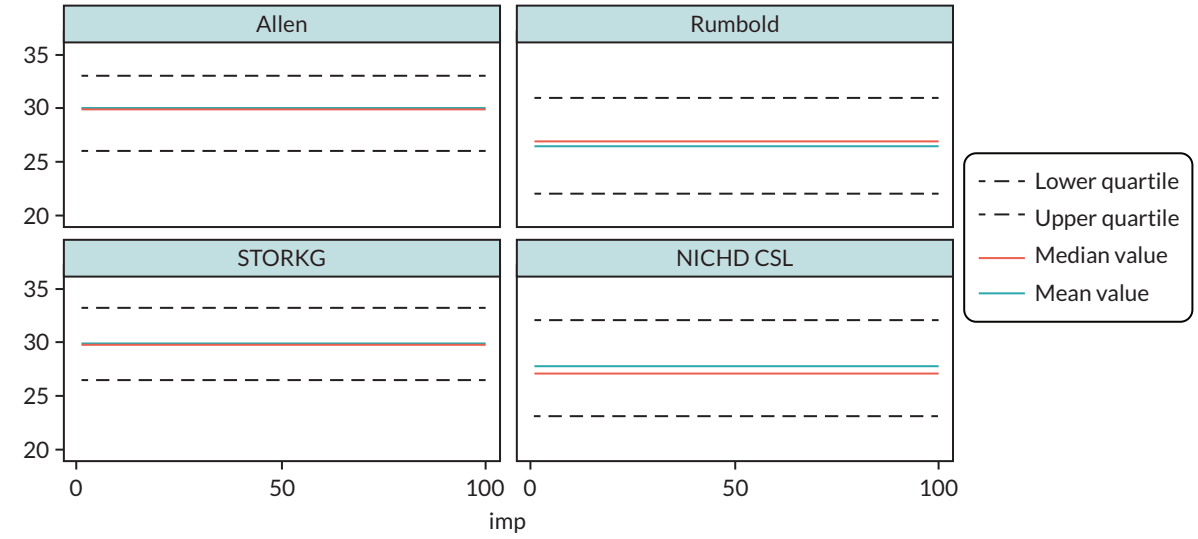

FIGURE 29 Mother's age.

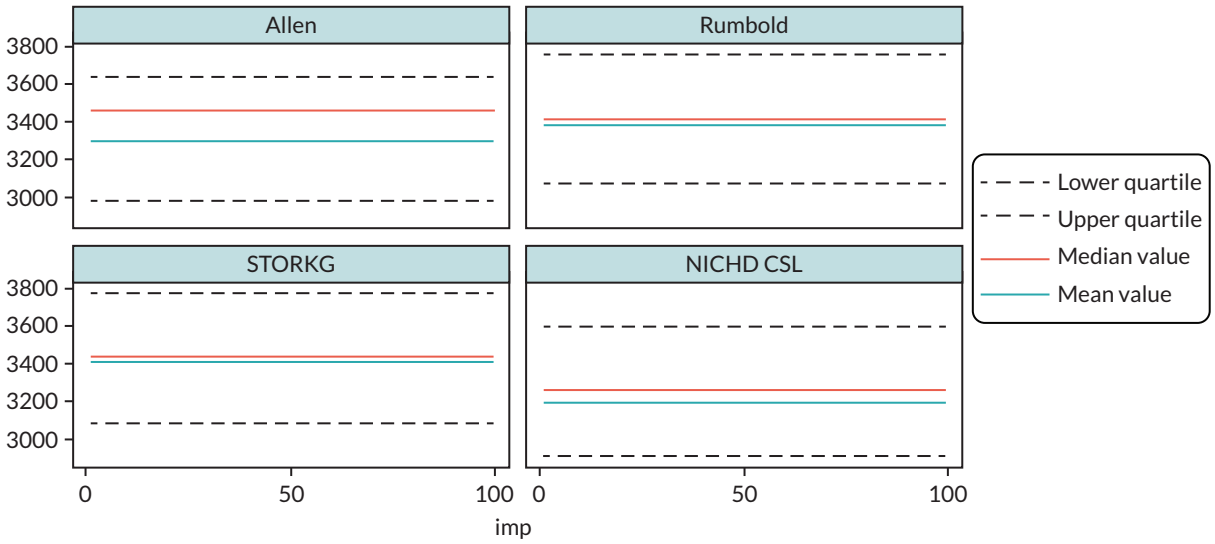

FIGURE 30 Birthweight.

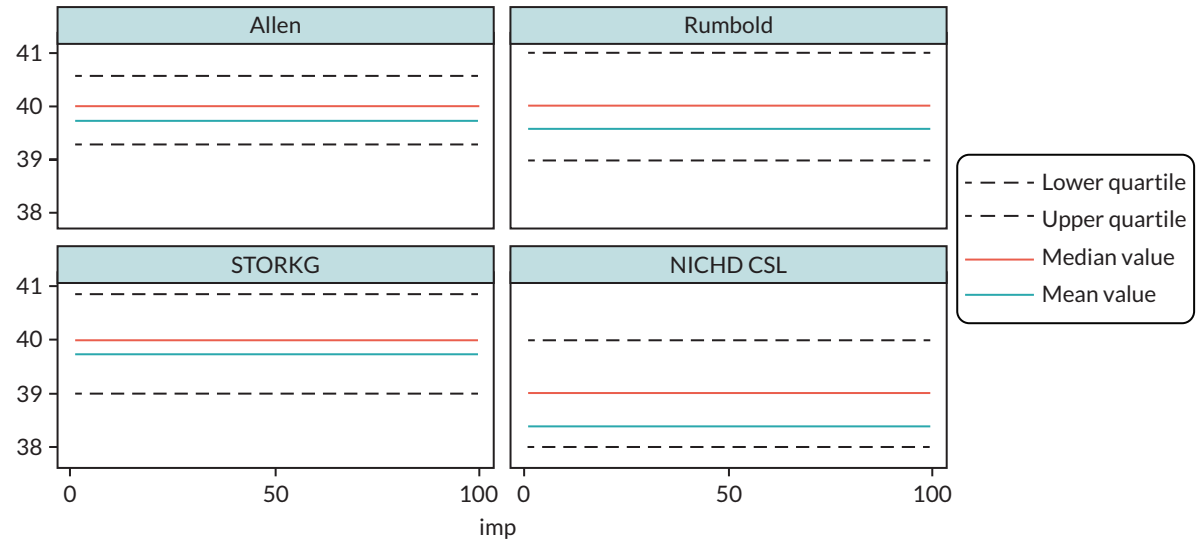

FIGURE 31 Gestational age at delivery.

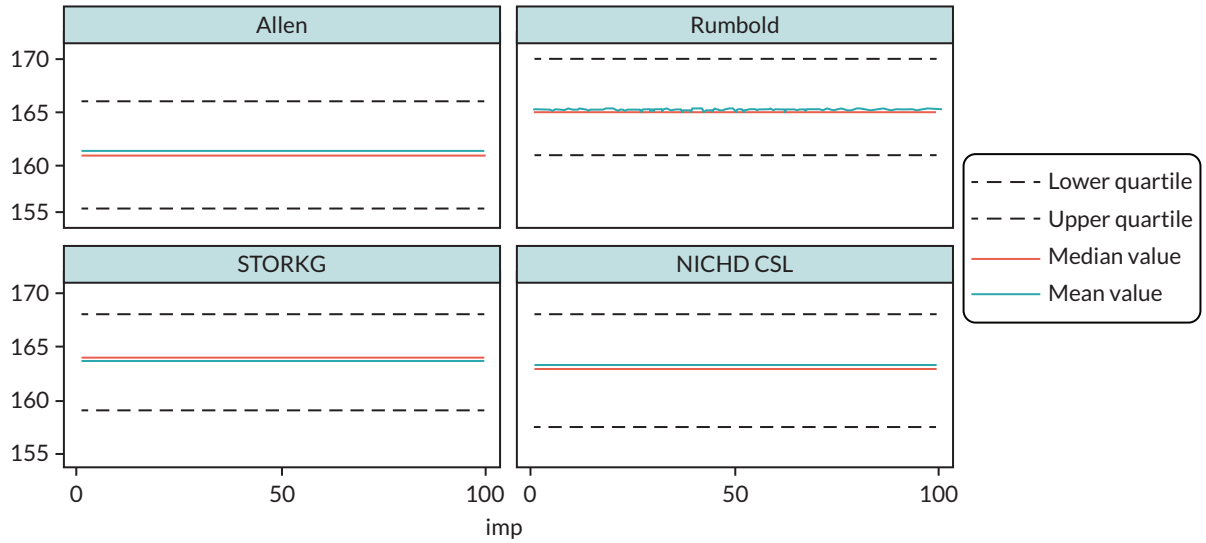

FIGURE 32 Mother's height.

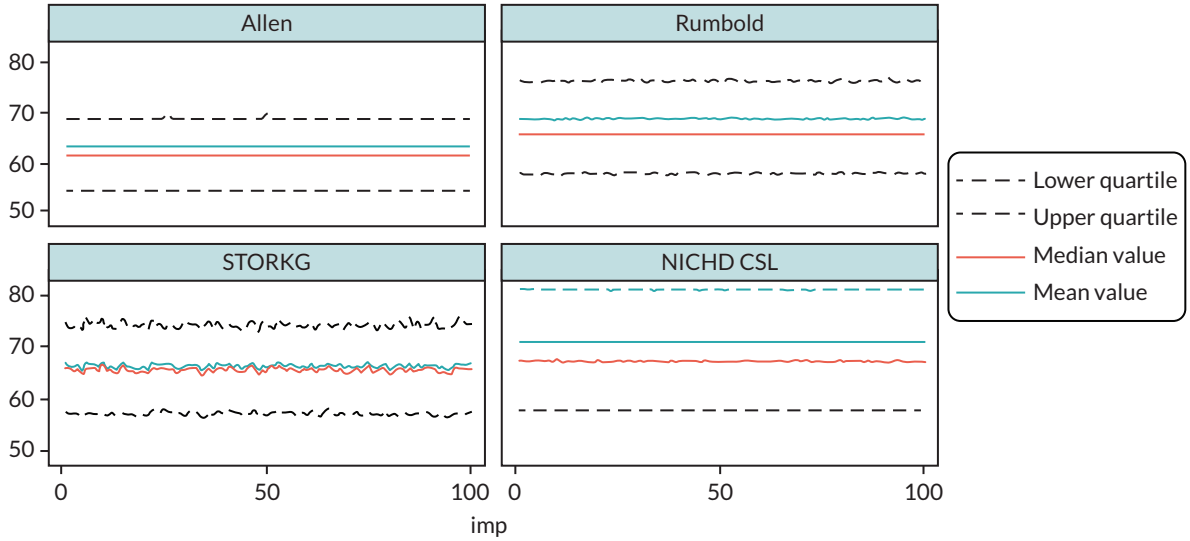

FIGURE 33 Mother's weight.

Appendix 6 Calculation of probabilities and cost values

TABLE 21 Calculation of probabilities and cost values

Parameter	Source	Description
CS	Hospital Episode Statistics 2018–9, Method of Delivery, table 3a	Divided # CS by # births excluding unknowns
Non-CS	Hospital Episode Statistics 2018–9, Method of Delivery, table 3a	1-P (CS)
CS (FGR)	Assumption	
FGR	Vieira 2019, NICE guidelines	10% of all babies are SGA, and 1/3 of them are FGR
Sensitivity of ultrasound scan	Haragan 2015	USAC < 5th percentile to predict BW < 10 percentile
Sensitivity of SFH measurement	Pay 2015	Accuracy of SF height for the prediction of SGA defined as BW ≥ 2 SDs below the mean
Sensitivity of (SFH + ultrasound)	Derived	Average of sensitivity of ultrasound and SFH measurement
Specificity of ultrasound scan	Haragan 2015	USAC < 5th percentile to predict BW < 10 percentile
Specificity of SFH measurement	Pay 2015	Accuracy of SF height for the prediction of SGA defined as BW ≥ 2 SDs below the mean
Specificity of (SFH + Ultrasound)	Derived	Average of specificity of ultrasound and SFH measurement
Sensitivity of prediction model	Prediction model, Section 6.3.3.2	Probability threshold 0.08
Specificity of prediction model	Prediction model, Section 6.3.3.2	Probability threshold 0.08
CS	NHS reference cost: national schedule of reference costs: the main schedule, total HRG's	Weighted average of planned (NZ50C) and emergency section (NZ51C)
Non-CS	NHS reference cost: national schedule of reference costs: the main schedule, total HRG's	Weighted average of NZ30C, NZ31C, NZ40C, NZ41C
SFH measurement	PSSRU 2006	
Ultrasound fetal growth scan	NHS reference cost: national schedule of reference costs: the main schedule, DADS	NZ73Z

EME
HSDR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library