NIHR | National Institute for Health and Care Research

Health Technology Assessment

Check for updates

Extended Research Article

Accuracy and clinical effectiveness of fetal growth monitoring strategies for the prediction of small for gestational age at birth: a systematic review and meta-analysis

Evangelos Danopoulos,[1] Nigel Armstrong,[1] Kevin McDermott,[1] Jiongyu Chen,[1] Xiaoyu Tian,[1] Caro Noake[1] and Marie Westwood[1,2*]

[1]Kleijnen Systematic Reviews (KSR) Ltd, York, UK
[2]Clinical Epidemiology and Public Health, Bristol Medical School, University of Bristol, UK

*Corresponding author marie.westwood@bristol.ac.uk

Published December 2025
DOI: 10.3310/AJLK7403

Scientific summary

Accuracy and clinical effectiveness of fetal growth monitoring strategies for the prediction of small for gestational age at birth: a systematic review and meta-analysis

Health Technology Assessment 2025; Vol. 29: No. 62
DOI: 10.3310/AJLK7403

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

Smallness for gestational age has been found to be associated with a higher risk of stillbirth. According to the latest report by the Office for National Statistics, low birthweight (BW) is one of the key risk factors for neonatal and infant mortality in England and Wales.

In the UK, pregnant people are typically screened using ultrasound (US) at around 12 weeks of gestation and approximately at the middle of pregnancy for gestational aging. After mid-pregnancy, they are offered further US monitoring, to assess fetal growth, only where clinically indicated.

The Healthcare Safety Investigation Branch (HSIB) 2021 National Learning Report, '*Intrapartum stillbirth: learning from maternity safety investigations that occurred during the COVID-19 pandemic, 1 April to 30 June 2020*', identified 11 cases of stillborn small for gestational age (SGA) infants, of whom 8 were not detected until birth, a finding which could be viewed as a further indication of the poor sensitivity of the methods currently used to detect SGA. The HSIB Report made eight safety recommendations, including Safety recommendation R/2021/148:

HSIB recommends that the Department of Health and Social Care commission a review to improve the reliability of existing assessment tools for fetal growth and fetal heart rate to minimise the risk for babies.

This systematic review has been commissioned and undertaken to summarise the available evidence to inform the fetal growth monitoring component of this recommendation.

Objectives

The overall objective of this project was to summarise the available evidence to inform the fetal growth monitoring component of the HSIB safety recommendation to '*commission a review to improve the reliability of existing assessment tools for fetal growth and fetal heart rate to minimise the risk for babies*'. The following research questions were defined to address the project objectives:

1. What are the effects, on clinical outcomes [e.g. neonatal morbidity, rates of brain injury, unplanned neonatal intensive care unit (NICU) admissions and parental morbidity] and rates of stillbirth and neonatal death, of interventions (e.g. iatrogenic delivery) which are made based on the findings of fetal growth monitoring to detect SGA/fetal growth restriction (FGR)?
    a. What are the effects of fetal growth monitoring to detect SGA/FGR on rates of pre-term iatrogenic delivery and gestational age at iatrogenic delivery?
2. What are the effects, on neonatal and parental outcomes, of implementing published guidelines for fetal growth monitoring?
3. What is the accuracy of different methods of monitoring fetal growth, for example, symphysis fundal height, US measurement of fetal size [fetal abdominal circumference (AC) or estimated fetal weight (EFW)], for predicting SGA/FGR at delivery?
4. What are the effects, on the performance (accuracy) of different methods of monitoring fetal growth, of key operational variables [e.g. timing of monitoring, type of growth reference chart used (customised or population-based), experience and training of clinical practitioner performing monitoring] and parental characteristics [e.g. body mass index (BMI)/obesity and ethnicity]?
    a. What factors affect the failure to obtain a satisfactory measurement or lack of clinical confidence in the reported result?

Methods

Assessment of clinical effectiveness

Nineteen databases, including MEDLINE and EMBASE, research registers, a conference proceedings resource and a pre-print resource were searched for relevant studies from 2000 to March 2023. The main EMBASE and MEDLINE searches were rerun in their entirety in September 2023. Search results were screened for relevance independently by two reviewers. Full-text inclusion assessment, data extraction and quality assessment were conducted by one reviewer and were checked by a second. The methodological quality of included diagnostic test accuracy (DTA) studies was assessed using Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2). The methodological quality of included randomised controlled trials (RCTs) was assessed using the revised Cochrane Risk of Bias Tool for Randomised Trials (RoB 2). The methodological quality of observational 'before-and-after' studies was assessed using a checklist, devised by the authors, for this review.

Where multiple accuracy studies assessed the same growth monitoring strategy, summary estimates of the sensitivity and specificity together with 95% confidence intervals (CIs) and prediction regions for the prediction of SGA at delivery were calculated. All meta-analyses involved fewer than four studies and estimated separate pooled estimates of sensitivity and specificity, using random-effects logistic regression. For comparative studies that evaluated the same clinical outcome(s), intervention/monitoring method and comparator, summary effect estimates with 95% CIs were calculated via random-effects models using the Mantel–Haenszel method with a Paule-Mandel estimator. Where between-study heterogeneity limited the applicability of meta-analysis, a narrative summary utilising text, tables and figures has been provided. The *Results of the assessment of clinical effectiveness* section of this report is structured by research question.

Results

Fifty-eight studies (78 publications) were included in the review.

Study quality

The methodological quality of diagnostic accuracy studies included in this review was high; in 44 of the 58 QUADAS-2 assessments, studies were rated as having low risk of bias (RoB) and low concerns regarding applicability on all domains.

Using RoB 2, all included RCTs were rated as having some concerns due to possible deviations from the intended interventions but were otherwise rated as low RoB.

For included comparative observational studies, the areas of concern were retrospective study design (three studies), inappropriate exclusion of participants and differences in selection criteria before and after the introduction of monitoring (one study).

Research question 1

Two studies (Michaeli J, Michaeli O, Rozitzky A, Grisaru-Granovsky S, Feldman N, Srebnik N. Application of prospect theory in obstetrics by evaluating mode of delivery and outcomes in neonates born small or appropriate for gestational age. *JAMA Network Open* 2022;**5**:e222177, and Nymark Hansen D, Sand Odgaard H, Uldbjerg N, Sinding M, Sorensen A. Screening for small for gestational-age fetuses. *Acta Obstet Gynecol Scand* 2019;**99**:503–9) provided data to inform research question 1. Both were retrospective cohort studies, which, in addition to providing some accuracy data for growth monitoring strategies using EFW US, also investigated the relationship between antenatal classification [fetal growth monitoring test result, SGA or appropriate for gestational age (AGA)], whether or not this classification was correct in relation to BW classification of SGA or AGA [test result category: true positive (TP), false positive (FP), false negative (FN) or true negative (TN)] and parental and neonatal clinical outcomes. Both studies reported some results indicating that both antenatal identification of SGA pregnancies (TP vs. FN) and antenatal misclassification of AGA pregnancies as SGA (FP vs. TN) were associated with increased rates of intervention (induction of labour and caesarean deliveries). Despite the apparent increased rates of intervention in SGA pregnancies that were identified antenatally, there was no evidence that the antenatal identification of SGA was associated with reduction in stillbirths and there was

no clear evidence of an association with improved neonatal outcomes; the only statistically significant effect identified was a reduction in the rate of NICU admission in SGA neonates identified antenatally (TP) compared with those not identified (FN), unadjusted odds ratio (OR), calculated from the data reported in Michaeli *et al.* (2022), 0.63 (0.52, 0.76). There was no evidence that antenatal misclassification of AGA neonates as SGA (FP) had any statistically significant detrimental effects on neonatal outcomes compared to correct classification as AGA (TN).

Research question 2

Four comparative observational studies and one RCT assessed the clinical effects of implementing guidance on fetal growth monitoring and all studies were concerned with the effects of implementing the Growth Assessment Protocol (GAP). The results of meta-analyses of three studies (n = 28,911) indicated that the GAP implementation was associated with an increase in the rates of induction of labour [pooled OR was 1.16 (95% CI 1.01 to 1.34), I^2 14%], but not in the risk of caesarean birth [pooled risk ratio (RR) was 1.05 (95% CI 0.86 to 1.27), I^2 61%]. The results of meta-analysis also indicated that the GAP implementation was associated with a reduction in the risk of stillbirth [RR 0.79 (95% CI 0.74 to 0.84), four studies, I^2 0%, n = 64,9272]. Additionally, in the only included RCT, the GAP implementation was associated with a small reduction in the rate of stillbirth, difference −0.07 (95% CI −0.14 to −0.01) (adjusted for age, ethnicity, parity and stratification factor). Only two studies reported data for the adverse neonatal outcomes, NICU admission and 5-minute Apgar score < 7; pooling of the results indicated that implementation was associated with a reduction in the risk of poor 5-minute Apgar score, RR of 0.78 (95% CI 0.64 to 0.95), I^2 0%, but it had no statistically significant effect on NICU admission, 0.59 (95% CI 0.02 to 20.03), I^2 93%. However, it should be noted that the adjusted difference in the rate of low Apgar scores reported in the RCT indicated that the GAP implementation had no statistically significant effect, −0.2 (95% CI −0.4 to 0.1).

Research question 3

Single US (assessing EFW US, AC US and combinations of these parameters) during the third trimester on singleton pregnancies were the most commonly evaluated growth monitoring strategies with respect to test accuracy. The majority of studies calculated EFW using a variation of the Hadlock equations. The most frequently used threshold for all the tests was < 10th percentile. In general, between-study comparisons of test accuracy are of limited value in that DTA studies are observational studies, and hence, comparing the performance of tests based on evaluations in different studies does not take account of differences in study population or other factors that may affect test performance. For this systematic review, between-study comparisons of the performance of different growth parameters were particularly problematic (even for the selected data set described above) due to the high degree of variation in other components of the test strategy (EFW equations, test reference charts and BW reference charts). Three studies, one conducted in a general unselected population and two conducted in high-risk populations, reported the results of within-study comparisons of EFW US calculated using Hadlock 2c, AC US and OR/AND combinations of the two parameters, using test and BW thresholds of < 10th percentile. The results of these studies consistently indicated that, as might be expected, the highest sensitivity was achieved when using a combination of EFW and AC, where the test threshold was defined as either parameter < 10th percentile (the OR combination), and that a combination where the test threshold was defined as both parameters < 10th percentile (the AND combination) gave the lowest sensitivity; EFW US alone consistently had either lower or equal sensitivity to AC US alone. The results of all of three within-study comparisons reported generally high specificity values (> 90%), with little variation in specificity across test strategies (EFW US calculated using Hadlock 2c, AC US and OR/AND combinations of the two parameters).

Comparison of different testing timings between the first, second and third trimesters indicated that the third-trimester testing strategies performed better with respect to maximising the antenatal detection of SGA babies. In terms of single versus multiple/serial testing, sensitivity estimates were very low for all strategies other than where multiple testing occurred during the third trimester. Both these outcomes suggest that there is more to be gained from testing during the third trimester alone; the limited available data do not support a benefit from earlier testing.

All studies, included in this review, that assessed the accuracy of growth monitoring strategies in unselected general populations, evaluated universal testing (screening) rather than clinically indicated testing as currently recommended in the UK. One UK study (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**:2089–97) compared the accuracy of the third-trimester EFW US

when applied universally in a general population to its accuracy when applied only where clinically indicated. This study used a general reference chart and a threshold of < 10th percentile for EFW and a local UK BW chart. For BW < 10th percentile, the sensitivity was 57% for universal testing and 20% for clinically indicated testing, with corresponding specificities of 90% and 98%.

Research question 4

There were insufficient data to adequately inform the assessment of how the accuracy of individual test strategies for fetal growth monitoring may vary with maternal characteristics (e.g. BMI/obesity and ethnicity) or the type, training and experience of the clinical practitioner performing the monitoring.

Most of the available data for research question 4 were concerned with variations in the accuracy of the metrics used to monitor fetal growth (EFW US and AC US) with the reference chart used for the index test or the reference chart used for BW. Nine included studies used 18 different reference charts for the index test and 15 different BW reference charts, rendering meaningful between-study comparisons challenging. Overall, no clear trend in test strategy performance was observable across studies, either for the use of general versus local reference charts (either EFW or BW) or for the use of non-sex-specific versus sex-specific BW reference charts. One large UK study (Mathewlynn S, Impey L, Ioannou C. Detection of small- and large-for-gestational age using different combinations of prenatal and postnatal charts. *Ultrasound Obstet Gynecol* 2022;**60**:373–80) used five different BW reference charts resulting in prevalences of SGA (< 10th percentile) at birth ranging from 4.2% to 10.9%. In Mathewlynn *et al.* (2022), the use of the locally derived reference chart for EFW always resulted in the highest sensitivity and lowest specificity estimates for any given BW reference chart (definition of SGA).

Conclusions

There is a lack of evidence linking fetal growth monitoring and the results of tests used to monitor fetal growth to changes in rates of stillbirth, perinatal death or adverse neonatal clinical outcomes.

There is some evidence to suggest that implementation of the GAP care pathway in UK settings may be associated with a reduction in adverse neonatal outcomes, but evidence about the effects of implementation on stillbirth was inconsistent. It is also unclear to what extent observed effects were attributable to GAP or other contemporaneous changes to routine care. The extent to which any effects of GAP, which is a complex, multifactorial intervention, may be attributable to the antenatal detection of SGA is also unknown.

There is a lack of evidence to assess the effectiveness of implementing UK clinical guidelines with respect to reducing rates of stillbirth and adverse neonatal outcomes.

Where fetal growth monitoring is to be implemented, there is insufficient evidence to support strongly favouring any one test strategy [combination of test timing, parameter measured (including any formula used, as for estimating EFW) and threshold and reference chart] for use in fetal growth monitoring. The available evidence suggests that testing during the third trimester is likely to result in a more accurate prediction of SGA at birth than earlier testing and that (for the general pregnant population) universal third-trimester US is likely to offer an improved sensitivity compared to testing based on clinical indication. Evidence from within-study comparisons of different test parameters suggests that, for a single US examination in the third trimester, a combination of EFW OR AC < 10th percentile could offer an increased sensitivity relative to either parameter used alone. There is also some evidence to suggest that, when using EFW US to monitor fetal growth, the use of a locally derived reference chart for EFW may result in the highest sensitivity for a given BW reference chart (definition of SGA).

Study registration

This study is registered as PROSPERO CRD42023408030.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Evidence Synthesis programme (NIHR award ref: NIHR135862) and is published in full in *Health Technology Assessment*; Vol. 29, No. 62. See the NIHR Funding and Awards website for further award information.

Health Technology Assessment

HTA TAR

ISSN 2046-4924 (Online)

Impact factor: 4

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 4 and is ranked 30th (out of 174 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 commissioned and funded by the Evidence Synthesis Programme on behalf of NICE as award number NIHR135862. The contractual start date was in March 2023. The draft manuscript began editorial review in November 2023 and was accepted for publication in January 2025. 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).