Quality control of ultrasound for fetal biometry: results from the INTERGROWTH-21st Project

Angelo Cavallaro[a], Stephen T Ash[a], Raffaele Napolitano[a], Sikolia Wanyonyi[b], Eric O Ohuma[a,c], Malid Molloholli[a], Joyce Sande[b], Ippokratis Sarris[a], Christos Ioannou[a], Tess Norris[a], Vera Donadono[a], Maria Carvalho[b], Manorama Purwar[e], Fernando C Barros[i,j], Yasmin A Jaffer[d], Enrico Bertino[f], Ruyan Pang[h], Michael G Gravett[g], Laurent J. Salomon[k], Julia Alison Noble[l], Douglas G Altman[c], Aris T Papageorghiou[a]

[a]*Nuffield Department of Obstetrics & Gynaecology and Oxford Maternal & Perinatal Health Institute, Green Templeton College, University of Oxford, Oxford, UK*

[b]*Faculty of Health Sciences, Aga Khan University, Nairobi, Kenya*

[c]*Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK*

[d]*Department of Family & Community Health, Ministry of Health, Muscat, Sultanate of Oman*

[e]*Nagpur INTERGROWTH-21st Research Centre, Ketkar Hospital, Nagpur, India*

[f]*Dipartimento di Scienze Pediatriche e dell'Adolescenza, Cattedra di Neonatologia, Università degli Studi di Torino, Torino, Italy*

[g]*Global Alliance to Prevent Prematurity and Stillbirth (GAPPS), Seattle, WA, USA*

[h]*School of Public Health, Peking University, Beijing, China*

[i]*Programa de Pós-Graduação em Saúde e Comportamento, Universidade Católica de Pelotas, Pelotas, RS, Brazil*

[j]*Programa de Pós-Graduação em Epidemiologia, Universidade Federal de Pelotas, Pelotas, RS, Brazil*

[k]*Maternité Necker-Enfants Malades, AP-HP, Université Paris Descartes, Paris, France*

[l]*Department of Engineering Science, University of Oxford, Oxford, UK*

Correspondence: Prof Aris Papageorghiou, Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Women's Centre, Level 3 John Radcliffe Hospital, Headington, Oxford OX3 9DU, UK

e-mail: aris.papageorghiou@obs-gyn.ox.ac.uk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.18811

Short title: Quality control in ultrasound fetal biometry

Key words: pregnancy; fetal growth; quality control; reproducibility; variability

ABSTRACT

Objectives: To assess a comprehensive package of ultrasound quality control in a large multicentre study of fetal growth – the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project.

Methods: We performed quality control (QC) measures on 20,313 ultrasound scan images taken prospectively from 4,321 fetuses at 14-41 weeks' gestation in eight geographical locations. At the time of each ultrasound examination, three fetal biometric variables were measured in triplicate on separately generated images: head circumference (HC), abdominal circumference (AC) and femur length (FL). All measurements were taken in a blinded fashion. QC had two elements: 1) qualitative QC: visual assessment by sonographers at each study site of their images based on specific criteria with 10% of images being re-assessed at the Oxford-based Ultrasound Quality Unit (compared using an adjusted kappa statistic), and 2) quantitative QC: measurement data were assessed by (a) comparing the first, second and third measurement (intraobserver variability); (b) re-measurement of caliper replacement in 10% (interobserver variability), both by Bland-Altman plots, and (c) plotting frequency histograms of the SDs of triplicate measurements and assessing how many were above or below 2SDs of the expected distribution. The system allowed the sonographers' performance to be regularly monitored.

Results: A high level of agreement between the self- and external scoring was demonstrated for all measurements (kappa = 0.99 [95% confidence interval: 0.98, 0.99]

for HC, 0.98 [0.97, 0.99] for AC, and 0.96 [0.95, 0.98] for FL. Intraobserver variability (95% limits of agreement (LoA)) of ultrasound measures for HC, AC and FL were ±3%, ±6% and ±6%, respectively; the corresponding values for interobserver variability were ±4%, ±6% and ±6%. The SD distribution of triplicate measurements for all biometric variables showed excessive variability for three of 31 sonographers, allowing prompt identification and retraining.

Conclusions: Qualitative and quantitative QC monitoring was feasible and highly reproducible in a large multicentre research study, which facilitated the production of high-quality ultrasound images. We recommend that the QC system we developed is implemented in future research studies and clinical practice.

INTRODUCTION

Standardisation and quality control (QC) of fetal ultrasound biometry are essential to ensure high levels of reproducibility among operators and ultrasound facilities. This particularly applies to multicentre studies because reproducibility and measurement consistency – among even well-trained sonographers – improve as a result of introducing QC systems [1]. Unfortunately, however, a common failing in this field is a complete absence of QC systems: for example, in studies designed to create charts for pregnancy dating, fetal and neonatal growth [2-5].

Although the effects of QC on measurement reproducibility have been demonstrated in research settings, their relevance may be even greater in routine clinical practice because measurement accuracy is critical for detecting abnormal fetal growth patterns, especially in the absence of blinding of measurements to the sonographer. In fact, avoiding false positive findings, with their attendant anxiety and risks of unnecessary interventions [6], is almost as important in antenatal care as diagnostic failures.

For a QC system in fetal biometry to be useful clinically, multiple strategies need to be employed [7], such as:

(i) Qualitative scoring of ultrasound images against predefined criteria [8];

(ii) Quantitative assessment of measurements and comparison with their expected distributions as, for example, occurs in fetal nuchal translucency QC [9-13] although, until now, these approaches have largely only been utilised in small studies [8-10, 13, 14].

Here we describe and assess the value of the comprehensive QC package used in the Fetal Growth Longitudinal Study (FGLS) of the INTERGROWTH-21st Project.

METHODS

Women at low risk of adverse pregnancy outcomes were recruited into FGLS, one of the three main components of INTERGROWTH-21st (www.intergrowth21.org.uk), a multicentre, multi-country, population-based project, conducted between 2008 and 2014 in eight countries [5, 15, 16], which aimed to construct international fetal growth standards. Serial ultrasound scans were performed every 5±1 weeks from 14+0 to 41+6 weeks' gestation. Gestational age was calculated on the basis of the last menstrual period (LMP) provided that: a) it was known and certain; b) the menstrual cycles were regular; c) there was no hormonal contraceptive use or breastfeeding during the 2 months prior to natural conception, and d) standardised [17] ultrasound measurement of the fetal crown-rump length between 9+0 and 13+6 weeks' gestation agreed with the LMP-based estimate of gestational age within 7 days [18].

At each examination, three fetal biometric variables were measured in triplicate on separately generated two-dimensional ultrasound images: head circumference (HC), abdominal circumference (AC) and femur length (FL). Thus, each examination produced nine measurements (three per variable) in accordance with the study protocol (www.intergrowth21.org.uk) [19].

All sonographers were recruited on the basis of being motivated, reliable and trained in ultrasound; ability to speak the local language(s) and work positively within a team structure. The goals of standardisation were, firstly to ensure that all sonographers fully understood the study protocol and take measurements in an identical fashion, and secondly that they were familiar with the equipment used. The precise details of how measurements were taken for FGLS and how data collection was standardised (through training, assessment and certification of all the sonographers) are presented in full elsewhere [7, 19]. Head measurements were obtained in the transthalamic plane, placing the calipers on the outer border of the skull, using both the ellipse facility and

two perpendicular diameters. Abdominal measurements were obtained in an axial plane, with the umbilical vein in the anterior third of the fetal abdomen (at the level of the portal sinus) and the stomach bubble visible. Again, both the ellipse facility and the two diameters method were used, placing the calipers on the outer border of the body outline (skin covering). In this study, we elected to analyse only the HC and AC measurements obtained using the ellipse facility, as a previous study showed that these were almost identical to those using the two diameters, but marginally more reproducible [20]. For FL, the femur closest to the probe was measured, with its long axis as horizontal as possible. Calipers were placed on the outer borders of the diaphysis of the femoral bone ('outer to outer').

All ultrasound scans were performed using the same commercially available ultrasound machine (Philips HD-9, Philips Ultrasound, Bothell, WA, USA) with curvilinear abdominal transducers (C5-2, C6-3 and V7-3). During the INTERGROWTH-21st Project blinding of operators to the measurement value was undertaken, thus eliminating expected value bias. For this purpose, the manufacturer programmed the machine's software so that the measurement values did not appear on screen during a scan.

The QC strategies adopted, which are described in detail below, included qualitative (i.e. image scoring) and quantitative analyses (i.e. estimating intraobserver and interobserver variability, and standard deviations (SD) of triplicate measures) for each biometric variable. Six sonographers undertook QC at the Oxford-based Ultrasound Quality Unit (USQU); any uncertainties were adjudicated by the QC Director (ATP). The analyses were performed monthly for the first 18 months of each site's participation and quarterly thereafter, or more frequently if any QC concerns were raised so as to identify sonographers performing outside accepted norms to allow corrective action (e.g. retraining) to be administered promptly [7].

Qualitative QC: Image scoring

Images were scored [7, 8] based on a set of criteria, each worth one point towards the total score, with a maximum of six points for HC and AC, and four points for FL (Table 1). All images were self-scored at the time of scanning by the sonographer taking the image. A randomly chosen sample of 10% of all these images was re-scored by a sonographer at the USQU; the highest of these three scores was used in the QC analysis.

In order to simplify the comparison between self- and USQU scoring, we divided data into low-scoring (1 to 3 for HC and AC, and 1 to 2 for FL) and high-scoring images (4 to 6 for HC and AC, and 3 to 4 for FL) [7]. As the quality was generally very good, higher scores were much more prevalent than lower scores; comparison between self- and USQU scoring was therefore undertaken using an adjusted kappa statistic (interobserver variability of image scoring) to account for the resulting unbalanced distributions of scores [21]. A kappa value of more than 0.6 was considered *a priori* an acceptable level of agreement among sonographers.

Quantitative QC: intraobserver and interobserver variation

As triplicate images and measurements were taken for each fetal biometric variable (HC, AC, FL), the *intraobserver* variability of the measurements could be assessed in the full dataset using Bland-Altman plots [22]. Instead of simply expressing differences within observers in actual measurement units (mm), pairwise comparisons were also made in percentage terms to account for changes in fetal size with increasing gestational age. The difference between two selected measurements was calculated and expressed as a percentage of their mean, then plotted against this mean. The 95% limits of agreement (LoA) were calculated and marked on the plots, giving a

quantifiable estimate for the measurement variability within the same observer associated with acquiring an image and positioning the calipers. These plots were generated by randomly selecting two of the three triplicate measurements taken at each scan for each biometric variable.

Actual (mm) and percentage difference Bland-Altman plots were also used to assess the *interobserver* variability of the measurements. As above, a sonographer at the USQU re-measured a random sample of 10% of all images from each site. The difference between the original and USQU measurements was expressed both as the actual value and as a percentage of their mean, then plotted against this mean. Again, 95% LoA were calculated and marked on the plots, giving a quantifiable estimate for the measurement reproducibility between observers associated with caliper placement.

Quantitative QC: data distribution

The SD of each measurement triplet was expressed as a percentage of the mean of the three measurements, enabling each sonographer's individual variability to be compared with the expected variability [14] whilst accounting for changes in fetal size with increasing gestational age.

Plotting each sonographer's SDs as separate frequency histograms allowed sonographers to be identified whose SD distributions differed from those of the expected range, based on the equivalent data derived from an initial variability study [14]. Sonographers demonstrating disproportionately large numbers of triplets outside the expected variability distribution were identified, causes investigated and retraining undertaken if necessary. Each sonographer's SDs were also plotted sequentially with a cumulative sum control chart [7, 14] to identify triplets with values >2 SDs more than 10% of the time.

All QC performed by USQU sonographers was undertaken blinded to the study site, sonographer identity, original measurement and their own repeated measurements. Unblinding only occurred to provide feedback to sonographers where necessary. All plots were generated and analyses performed using SAS software (Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA).

The INTERGROWTH-21st Project protocol was approved by the Oxfordshire Research Ethics Committee C (reference: 08/H0606/139); all the pregnant women enrolled gave informed written consent.

RESULTS

We studied 4,321 singleton fetuses, whose intrauterine biometric measures were used to construct international fetal growth standards, all of whom were born alive without congenital malformations [18]. Each fetus had a median of 5 (range 1-7; mean 4.9 [SD 0.8]; total 20,313) scans. Figure 1 shows the total number of scans eligible for QC for each biometric variable (HC, AC and FL) after excluding measures >5 SD's and missing data.

Between 20,040 and 20,313 scans, depending on the biometric variable, were assessed for intraobserver variability (Figure 1, Box 1). Ten per cent (n=1,735) of the 17,350 scans performed at sites other than Oxford were randomly selected for external image re-scoring and re-measurement by a USQU sonographer. Of these, 122 scans could not be assessed (due to incomplete backup of images or data lost due to corruption of the backup file) leaving 1,613 scans from 1,322 women that underwent QC by USQU sonographers (Figure 1).

Qualitative QC: Image scoring

Of the 1,613 scans, 1,340 (83.1%) were re-scored; the remainder were missing original image scores (n=256) or re-scorings were not logged (n=17). Overall, the quality of all measurements was high. The median self-scored image values for HC, AC and FL were 6 [interquartile range (IQR): 6-6], 6 [IQR: 6-6] and 4 [IQR: 4-4], respectively – the maximum values in the scoring system. There was a very high level of agreement between the self- and external scoring of image quality for all measurements, with adjusted kappa values of 0.99 [95% confidence interval (CI): 0.98, 0.99] for HC, 0.98 [95% CI: 0.97, 0.99] for AC, and 0.96 [95% CI: 0.95, 0.98] for FL (Table 2). In almost all

cases, both the local and USQU sonographers classified the same image as high-scoring (99% scored 4 to 6 for HC and AC, 98% scored 3 to 4 for FL).

This external image assessment process resulted in six sonographers requiring retraining over the entire study period, after which improvements in performance were seen.

Quantitative QC: intraobserver and interobserver variation

Intraobserver variability was assessed using all 20,313 scans, comparing two of the triplicate measurements for each scan, selected randomly and in random order. Overall, the reproducibility was very good (Supplementary figure 1). For HC, the mean difference was 0.0% and the 95% LoA were consistently between ±3.1 and ±3.5%. For AC, the mean difference was approximately 0.1% and the 95% LoA were consistently between ±5.4 and ±6.0%. For FL, the mean difference was approximately 0.2%; however, the 95% LoA, even when expressed as a percentage of FL, varied with gestation, showing greater variability and poorer intraobserver reproducibility at lower gestational ages. On average, the 95% LoA for FL were between ±5.8 and ±6.4% (Table 3).

Interobserver variability was assessed using 1,483 of the 1,613 scans selected in the 10% QC sample as 130 re-measured scans were erased as a result of a technical problem. Overall, the reproducibility was very good (Supplementary figure 2). For HC, AC and FL the mean difference was 1.0% or less, with 95% LoA between ±4.3 and ±4.4%, between ±5.9 and ±6.3% and between ±5.4 and ±5.9%, respectively (Table 3).

Quantitative QC: data distribution

Comparisons with the expected distribution of measures showed no cause for concern in any biometric variable for 28 out of the 31 study sonographers. In one instance, a sonographer was found to have 16.5% of HC SDs and 11.1% of AC SDs outside the expected range, whilst two other sonographers demonstrated unacceptably high FL SDs (Figure 2). In all three instances retraining was undertaken and improvements were seen thereafter. The total number of images taken by these three sonographers made up only a very small proportion of the total dataset [HC: 188 of 20,041 (0.9%); AC: 45 of 20,135 (0.2%); FL: 267 of 20,313 (1.3%)].

DISCUSSION

We report the implementation and results of using a comprehensive system to assess the quality of ultrasound images obtained from a large multicentre, international project. We not only demonstrate the system's feasibility, but show that it is possible to achieve a high level of reproducibility in such a study with the necessary QC measures.

Firstly, over 98% of the scored images were considered as high quality by both the local and USQU sonographers (qualitative QC); secondly, the intra- and interobserver reproducibility of measurements (quantitative QC) was high and within the limits of a previous study [14] (Table 3); and thirdly, we monitored images and data regularly, which enabled us to identify a few sonographers whose performance fell outside expected standards, following which corrective action was taken. It should be noted that this entire process relied upon initial training and standardisation [1, 7, 19] – a crucial element of the project's success.

Meticulous standardisation and ongoing monitoring of adherence to measurement protocols during data collection have been shown to ensure consistency and minimise systematic error in multicentre studies [1, 7, 11, 17]. In two recent systematic reviews of the literature relating to the creation of fetal crown-rump length charts and growth charts, no studies reporting a comprehensive QC process were identified, which undoubtedly contributed to the poor quality of many existing studies [2, 3].

Our study has a number of strengths: a) the QC strategy was prospectively designed and implemented [7, 15, 19], and based firmly on previous studies that assessed the role of feedback on image quality [7, 10, 23, 24], and b) visual assessment of ultrasound images was based on an objective criterion-based scoring system, which has been shown to be significantly more reproducible than subjective methods [8, 25, 26]. In Salomon et al.'s [8] original description of this process, high reproducibility levels for the image scoring

method were demonstrated (kappa between 0.60 and 0.98); despite undertaking QC in a blinded fashion, the results from our study were even better. The high level of reproducibility of such objective methods [9, 12, 25, 26] are corroborated by our study, which is the largest to date. It is likely that the high level of training of the sonographers acquiring the images and those conducting the QC, and the requirement for standardisation of all staff in settings of near-optimal conditions for scanning, contributed to the overall quality.

One of the limitations of the study is that only 10% of images underwent external scoring. However, this is the largest quality control programme ever performed in the setting of a study into fetal growth. All 100% images underwent self-scoring, and those images that were externally scored were randomly selected, meaning that there was no evidence that a different proportion would have yielded different results. Of course, implementing such a QC strategy is labour-intensive. While it is relatively easy to assess data distributions routinely [27], external qualitative assessment using image scoring requires additional resources.

More cost-effective options might include:

- Voluntary submission of a small number of selected images (as, for example, in certification for nuchal translucency measurement) [28]; however, the small number of images and the nature of self-selection mean it is difficult to ascertain whether such images are truly representative of a sonographer's routine practice.
- Self-assessment of images which correlates well with external scoring using a 10% random sample of all images suggesting that it may be a reasonable alternative. However, there are three reasons to be cautious: firstly, we have demonstrated that self-scoring is effective *in association with* external scrutiny, and it is not known whether similarly high quality is achievable without a QC

system; secondly, such a system is feasible with a few highly trained and motivated sonographers and may not be scalable, for example, to a national screening programme; and thirdly, while we have demonstrated excellent agreement between self- and external scoring across the whole dataset, the role of QC in screening is exception reporting, i.e. to detect individual outliers, rather than to demonstrate that, *on average,* the system works. Only by integrating all the elements of our QC system were we able to identify opportunities for improvement that could not be detected by self-scoring alone.

- Automated methods for QC of routinely collected images are being studied and may, in the future, be the best option. These systems have the potential advantage of allowing all images to be assessed objectively and at low cost [29, 30].

Regarding quantitative QC, a literature search was performed to identify previous publications on the evaluation of reproducibility of fetal ultrasound biometry after 14+0 weeks (17 studies identified) [1, 14, 27, 31-45]. Studies were selected only if reliable quantitative values were calculated as LoA or repeatability coefficients [46, 47]. Overall intraobserver reproducibility reported 95% LoA of less than 4% (12 mm) for HC, 6% (12 mm) for AC, and 7% (3 mm) for FL [1, 14, 32-37, 40, 41, 43, 45]. Similarly, for interobserver analyses 95% LoA for HC, AC and FL were within 4%, 6%, and 6%, respectively [1, 14, 27, 31, 34-39, 42, 44, 45]. Even though these studies were undertaken on smaller numbers of cases, mostly in single centre research settings and without blinding of the measurements, these values are not markedly different from the results of our large-scale multicentre study.

Our study has shown that, in general, both intra- and interobserver variability remain reasonably constant throughout pregnancy when reported as a percentage of fetal size. The exception is for FL, where increased variance was demonstrated at early

gestational ages, most likely due to the difficulty in accurately measuring FL when it is only 10-30 mm long.

In conclusion, both qualitative and quantitative QC monitoring were found to be feasible in a large multicentre fetal growth study. The development of a standardised fetal biometric ultrasound measurement protocol, standardisation of all sonographers (involving their training, assessment and certification), consistency and blinding of measurement are all necessary to minimise systematic error and ensure high reproducibility. Having developed a framework for ultrasound QC, we recommend that it is implemented in future similar research studies and, ideally, in clinical practice.

Conflict of interest: ATP is the Chief Medical Officer of Intelligent Ultrasound and receives non-financial support from Philips Ultrasound. JAN has received personal fees from Intelligent Ultrasound and grants and non-financial support from Philips Ultrasound. All other authors declare no competing interests.

Acknowledgement: This project was supported by a generous grant from the Bill & Melinda Gates Foundation to the University of Oxford (Oxford, UK), for which we are very grateful. We also thank the Health Authorities in Pelotas, Brazil; Beijing, China; Nagpur, India; Turin, Italy; Nairobi, Kenya; Muscat, Oman; Oxford, UK; and Seattle, WA, USA who facilitated the project by allowing participation of these study sites as collaborating centres. We are extremely grateful to Philips Medical Systems who provided the ultrasound equipment and technical assistance throughout the project.

REFERENCES

1. Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, Shah J, Sohoni S, Al Zidjali W, Hoch L, Altman DG, Papageorghiou AT. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. *Ultrasound Obstet Gynecol* 2011; **38**: 681-687.

2. Napolitano R, Dhami J, Ohuma EO, Ioannou C, Conde-Agudelo A, Kennedy SH, Villar J, Papageorghiou AT. Pregnancy dating by fetal crown-rump length: a systematic review of charts. *BJOG* 2014; **121**: 556-565.

3. Ioannou C, Talbot K, Ohuma E, Sarris I, Villar J, Conde-Agudelo A, Papageorghiou AT. Systematic review of methodology used in ultrasound studies aimed at creating charts of fetal size. *BJOG* 2012; **119**: 1425-1439.

4. Villar J, Giuliani F, Bhutta ZA, Bertino E, Ohuma EO, Ismail LC, Barros FC, Altman DG, Victora C, Noble JA, Gravett MG, Purwar M, Pang R, Lambert A, Papageorghiou AT, Ochieng R, Jaffer YA, Kennedy SH. Postnatal growth standards for preterm infants: the Preterm Postnatal Follow-up Study of the INTERGROWTH-21(st) Project. *Lancet Glob Health* 2015; **3**: e681-691.

5. Villar J, Cheikh Ismail L, Victora CG, Ohuma EO, Bertino E, Altman DG, Lambert A, Papageorghiou AT, Carvalho M, Jaffer YA, Gravett MG, Purwar M, Frederick IO, Noble AJ, Pang R, Barros FC, Chumlea C, Bhutta ZA, Kennedy SH. International standards for newborn weight, length, and head circumference by gestational age and sex: the Newborn Cross-Sectional Study of the INTERGROWTH-21st Project. *Lancet* 2014; **384**: 857-868.

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

7. Sarris I, Ioannou C, Ohuma EO, Altman DG, Hoch L, Cosgrove C, Fathima S, Salomon LJ, Papageorghiou AT. Standardisation and quality control of ultrasound

measurements taken in the INTERGROWTH-21st Project. *BJOG* 2013; **120 Suppl 2**: 33-37.

8. Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. *Ultrasound Obstet Gynecol* 2006; **27**: 34-40.

9. Fries N, Althuser M, Fontanges M, Talmant C, Jouk PS, Tindel M, Duyme M. Quality control of an image-scoring method for nuchal translucency ultrasonography. *Am J Obstet Gynecol* 2007; **196**: 272.e1-272.e5.

10. Chalouhi GE, Salomon LJ, Fontanges M, Althuser M, Haddad G, Scemama O, Chabot JM, Duyme M, Fries N. Formative assessment based on an audit and feedback improves nuchal translucency ultrasound image quality. *J Ultrasound Med* 2013; **32**: 1601-1605.

11. Snijders RJ, Thom EA, Zachary JM, Platt LD, Greene N, Jackson LG, Sabbagha RE, Filkins K, Silver RK, Hogge WA, Ginsberg NA, Beverly S, Morgan P, Blum K, Chilis P, Hill LM, Hecker J, Wapner RJ. First-trimester trisomy screening: nuchal translucency measurement training and quality assurance to correct and unify technique. *Ultrasound Obstet Gynecol* 2002; **19**: 353-359.

12. Herman A, Maymon R, Dreazen E, Caspi E, Bukovsky I, Weinraub Z. Nuchal translucency audit: a novel image-scoring method. *Ultrasound Obstet Gynecol* 1998; **12**: 398-403.

13. Gabriel CC, Echevarria M, Rodriguez I, Serra B. Analysis of quality of nuchal translucency measurements: its role in prenatal diagnosis. *ScientificWorldJournal* 2012; **2012**: 482832.

14. Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, Altman DG, Papageorghiou AT. Intra- and interobserver variability in fetal ultrasound measurements. *Ultrasound Obstet Gynecol* 2012; **39**: 266-273.

15. Villar J, Altman DG, Purwar M, Noble JA, Knight HE, Ruyan P, Cheikh Ismail L, Barros FC, Lambert A, Papageorghiou AT, Carvalho M, Jaffer YA, Bertino E, Gravett MG, Bhutta ZA, Kennedy SH. The objectives, design and implementation of the INTERGROWTH-21st Project. *BJOG* 2013; **120 Suppl 2**: 9-26.

16. Villar J, Papageorghiou AT, Pang R, Ohuma EO, Cheikh Ismail L, Barros FC, Lambert A, Carvalho M, Jaffer YA, Bertino E, Gravett MG, Altman DG, Purwar M, Frederick IO, Noble JA, Victora CG, Bhutta ZA, Kennedy SH. The likeness of fetal growth and newborn size across non-isolated populations in the INTERGROWTH-21st Project: the Fetal Growth Longitudinal Study and Newborn Cross-Sectional Study. *Lancet Diabetes Endocrinol* 2014; **2**: 781-792.

17. Ioannou C, Sarris I, Hoch L, Salomon LJ, Papageorghiou AT. Standardisation of crown-rump length measurement. *BJOG* 2013; **120 Suppl 2**: 38-41.

18. Papageorghiou AT, Ohuma EO, Altman DG, Todros T, Cheikh Ismail L, Lambert A, Jaffer YA, Bertino E, Gravett MG, Purwar M, Noble JA, Pang R, Victora CG, Barros FC, Carvalho M, Salomon LJ, Bhutta ZA, Kennedy SH, Villar J. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *Lancet* 2014; **384**: 869-879.

19. Papageorghiou AT, Sarris I, Ioannou C, Todros T, Carvalho M, Pilu G, Salomon LJ. Ultrasound methodology used to construct the fetal growth standards in the INTERGROWTH-21st Project. *BJOG* 2013; **120 Suppl 2**: 27-32.

20. Napolitano R, Donadono V, Ohuma EO, Knight CL, Wanyonyi SZ, Kemp B, Norris T, Papageorghiou AT. Scientific basis for standardization of fetal head measurements by ultrasound: a reproducibility study. *Ultrasound Obstet Gynecol* 2016; **48**: 80-85.

21. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993; **46**: 423-429.

22. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol* 2003; **22**: 85-93.

23. Jaudi S, Granger B, Herpin CN, Fries N, Du Montcel ST, Dommergues M. Online audit and feedback improve fetal second-trimester four-chamber view images: a randomised controlled trial. *Prenat Diagn* 2013; **33**: 959-964.

24. Thia EW, Wei X, Tan DT, Lai XH, Zhang XJ, Oo SY, Yeo GS. Evaluation of an objective method of image assessment for first-trimester nasal bone. *Ultrasound Obstet Gynecol* 2011; **38**: 533-537.

25. McLennan A, Schluter PJ, Pincham V, Hyett J. First-trimester fetal nasal bone audit: evaluation of a novel method of image assessment. *Ultrasound Obstet Gynecol* 2009; **34**: 623-628.

26. Wanyonyi SZ, Napolitano R, Ohuma EO, Salomon LJ, Papageorghiou AT. Image-scoring system for crown-rump length measurement. *Ultrasound Obstet Gynecol* 2014; **44**: 649-654.

27. Rijken MJ, Lee SJ, Boel ME, Papageorghiou AT, Visser GH, Dwell SL, Kennedy SH, Singhasivanon P, White NJ, Nosten F, McGready R. Obstetric ultrasound scanning by local health workers in a refugee camp on the Thai-Burmese border. *Ultrasound Obstet Gynecol* 2009; **34**: 395-403.

28. D'Alton ME, Cleary-Goldman J, Lambert-Messerlian G, Ball RH, Nyberg DA, Comstock CH, Bukowski R, Berkowitz RL, Dar P, Dugoff L, Craigo SD, Timor IE, Carr SR, Wolfe HM, Dukes K, Canick JA, Malone FD. Maintaining quality assurance for sonographic nuchal translucency measurement: lessons from the FASTER Trial. *Ultrasound Obstet Gynecol* 2009; **33**: 142-146.

29. Maraci MA, Bridge CP, Napolitano R, Papageorghiou A, Noble JA. A framework for analysis of linear ultrasound videos to detect fetal presentation and heartbeat. *Med Image Anal* 2017; **37**: 22-36.

30. Sarris I, Ohuma E, Ioannou C, Sande J, Altman DG, Papageorghiou AT. Fetal biometry: how well can offline measurements from three-dimensional volumes substitute real-time two-dimensional measurements? *Ultrasound Obstet Gynecol* 2013; **42**: 560-570.

31. Chang TC, Robson SC, Spencer JA, Gallivan S. Ultrasonic fetal weight estimation: analysis of inter- and intra-observer variability. *J Clin Ultrasound* 1993; **21**: 515-519.

32. Dudley NJ, Chapman E. The importance of quality management in fetal measurement. *Ultrasound Obstet Gynecol* 2002; **19**: 190-196.

33. Neufeld LM, Wagatsuma Y, Hussain R, Begum M, Frongillo EA. Measurement error for ultrasound fetal biometry performed by paramedics in rural Bangladesh. *Ultrasound Obstet Gynecol* 2009; **34**: 387-394.

34. Perni SC, Chervenak FA, Kalish RB, Magherini-Rothe S, Predanic M, Streltzoff J, Skupski DW. Intraobserver and interobserver reproducibility of fetal biometry. *Ultrasound Obstet Gynecol* 2004; **24**: 654-658.

35. Johnsen SL, Wilsgaard T, Rasmussen S, Sollien R, Kiserud T. Longitudinal reference charts for growth of the fetal head, abdomen and femur. *Eur J Obstet Gynecol Reprod Biol* 2006; **127**: 172-185.

36. Verburg BO, Steegers EA, De Ridder M, Snijders RJ, Smith E, Hofman A, Moll HA, Jaddoe VW, Witteman JC. New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study. *Ultrasound Obstet Gynecol* 2008; **31**: 388-396.

37. Pang MW, Leung TN, Sahota DS, Lau TK, Chang AM. Customizing fetal biometric charts. *Ultrasound Obstet Gynecol* 2003; **22**: 271-276.

38. Exacoustos C, Rosati P, Rizzo G, Arduini D. Ultrasound measurements of fetal limb bones. *Ultrasound Obstet Gynecol* 1991; **1**: 325-330.

39. Di Battista E, Bertino E, Benso L, Fabris C, Aicardi G, Pagliano M, Bossi A, De Biasio P, Milani S. Longitudinal distance standards of fetal growth. Intrauterine and Infant Longitudinal Growth Study: IILGS. *Acta Obstet Gynecol Scand* 2000; **79**: 165-173.

40. Hadlock FP, Deter RL, Harrist RB, Park SK. Fetal head circumference: relation to menstrual age. *AJR Am J Roentgenol* 1982; **138**: 649-653.

41. Hadlock FP, Deter RL, Harrist RB, Park SK. Fetal abdominal circumference as a predictor of menstrual age. *AJR Am J Roentgenol* 1982; **139**: 367-370.

42. Hadlock FP, Harrist RB, Deter RL, Park SK. Fetal femur length as a predictor of menstrual age: sonographically measured. *AJR Am J Roentgenol* 1982; **138**: 875-878.

43. Larsen T, Petersen S, Greisen G, Larsen JF. Normal fetal growth evaluated by longitudinal ultrasound examinations. *Early Hum Dev* 1990; **24**: 37-45.

44. Sarmandal P, Bailey SM, Grant JM. A comparison of three methods of assessing inter-observer variation applied to ultrasonic fetal measurement in the third trimester. *Br J Obstet Gynaecol* 1989; **96**: 1261-1265.

45. Al-Meshari AA RH, Raber H. Fetal biparietal diameter in Saudi Arabia. *Ann Saudi Med* 1987; **7**: 227-233.

46. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307-310.

47. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304**: 1491-1494.

TABLES

Table 1 - Image scoring criteria used for standardisation and quality control.

Cephalic plane (max. 6 points)	**Abdominal plane (max. 6 points)**	**Femoral plane (max. 4 points)**
1 Symmetrical plane	1 Symmetrical plane	1 Both ends of the bone clearly visible
2 Thalami visible	2 Stomach bubble visible	2 Angle <45°
3 Cavum septi pellucidi visible	3 Portal sinus visible	3 Femur occupying at least 30% of image
4 Cerebellum not visible	4 Kidneys not visible	4 Callipers placed correctly
5 Head occupying at least 30% of image	5 Abdomen occupying at least 30% of image	-
6 Callipers/ellipse placed correctly	6 Callipers/ellipse placed correctly	-

Table 2 - Matrix showing the number of scans that were self-scored (rows) versus external scoring results (columns) for head circumference, abdominal circumference and femur length, and adjusted kappa values.

			Quality score: External scoring					Adjusted kappa (95% CI)
			1 - 3	**4**	**5**	**6**	**Total**	
Quality score: Self scoring	**Head circumference**	**1 - 3**	**1**	1	4	2	**8**	0.99 (0.98 to 0.99)
		4	0	**3**	6	23	**32**	
		5	1	4	**26**	182	**213**	
		6	1	15	137	**917**	**1070**	
		Total	**3**	**23**	**173**	**1124**	**1323**	
	Abdominal circumference	**1 - 3**	**0**	0	0	7	**7**	0.98 (0.97 to 0.99)
		4	1	**3**	6	36	**46**	
		5	1	2	**20**	234	**257**	
		6	1	10	90	**906**	**1007**	
		Total	**3**	**15**	**116**	**1183**	**1317**	
			1	**2**	**3**	**4**		
	Femur length	**1**	**0**	0	0	0	**0**	0.96 (0.95 to 0.98)
		2	0	**0**	2	0	**2**	
		3	0	1	**7**	19	**27**	
		4	0	21	136	**1145**	**1302**	
		Total	**0**	**22**	**145**	**1164**	**1331**	

Table 3 - Quantitative QC: intraobserver variability for image acquisition and caliper placement and interobserver variability for caliper replacement (10% of all images), expressed as percentages (%). The relevant data from a previous study [14] are included here for comparison. QC: quality control; HC: head circumference; AC: abdominal circumference; FL: femur length.

		Quantitative QC			
		Pilot study		Our study	
		Mean difference	95% limits of agreement	Mean difference	95% limits of agreement
Intraobserver reproducibility (%) for image acquisition and caliper placement	**HC**	0.0	± 3.0	0.0	± 3.3
	AC	- 0.3	± 5.3	0.0	± 5.6
	FL	0.2	± 6.6	0.0	± 6.2
Interobserver reproducibility (%) for caliper replacement (10% of all images)	**HC**	1.3	± 3.7	1.0	± 4.4
	AC	1.1	± 5.7	- 0.1	± 6.0
	FL	0.8	± 5.8	- 0.8	± 5.6

FIGURE LEGENDS

Figure 1 - Flow chart of patients and scans included in the analysis. QC: quality control; HC: head circumference; AC: abdominal circumference; FL: femur length.

Figure 2 - Distribution of standard deviations (SD) (expressed as a percentage of the mean) of triplicate measurements of head circumference made by two sonographers. The vertical lines indicate the 97.5th centile value (2.42), median and 2.5th centile (0.16) taken from the reference standard study [7]. We illustrate two examples: in the upper panel 3% of triplicate measurements are above the 97.5th centile. In the lower panel 17% of measurements were above the accepted threshold, set at 10%; and retraining was undertaken.

Supplementary figure 1 - Bland Altman plots: intraobserver variability in head circumference (a,b), abdominal circumference (c,d) and femur length (e,f) measurements, expressed as millimetres (mm) (a,c,e) and percentage (%) (b,d,f).

Supplementary figure 2 - Bland Altman plots: interobserver variability in head circumference (a,b), abdominal circumference (c,d) and femur length (e,f) measurements, expressed as millimetres (mm) (a,c,e) and percentage (%) (b,d,f).

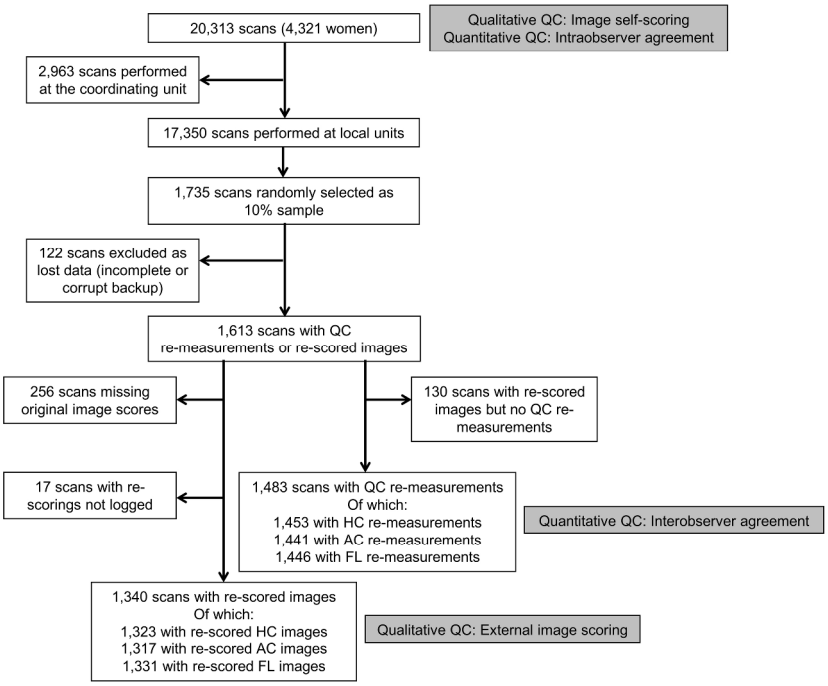

Figure 1 - Flow chart of patients and scans included in the analysis. QC: quality control; HC: head circumference; AC: abdominal circumference; FL: femur length.

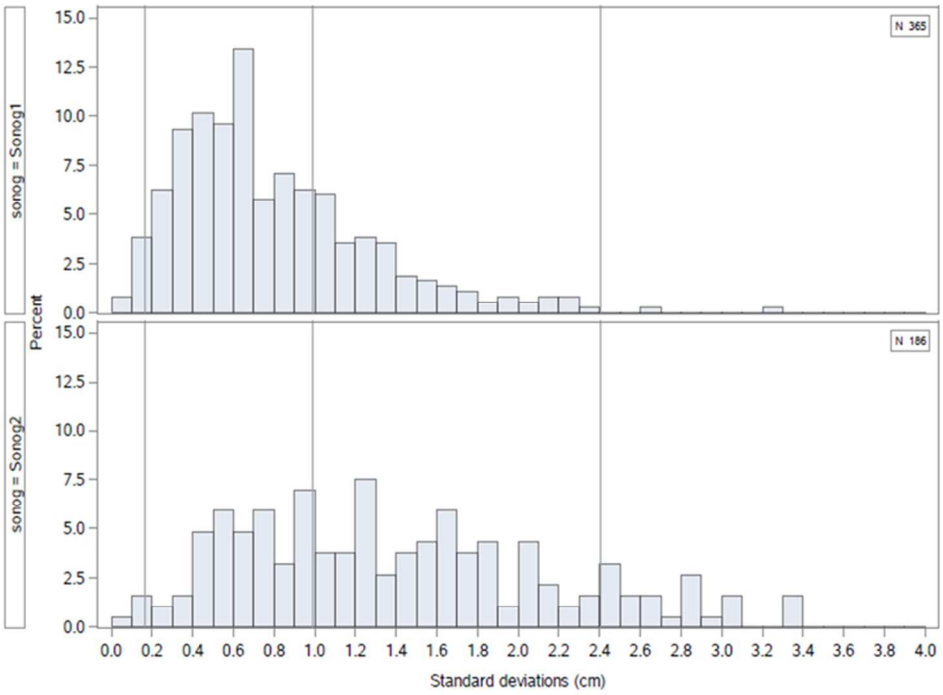

Figure 2 - Distribution of standard deviations (SD) (expressed as a percentage of the mean) of triplicate measurements of head circumference made by two sonographers. The vertical lines indicate the 97.5th centile value (2.42), median and 2.5th centile (0.16) taken from the reference standard study[7]. We illustrate two examples: in the upper panel 3% of triplicate measurements are above the 97.5th centile. In the lower panel 17% of measurements were above the accepted threshold, set at 10%; and retraining was undertaken.