Published in final edited form as:
J Ultrasound Med. 2016 August ; 35(8): 1725–1733. doi:10.7863/ultra.15.09087.

Ultrasound Quality Assurance (QA) for Singletons in the NICHD Fetal Growth Studies

Mary L. Hediger, PhD,
Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD

Karin M. Fuchs, MD,
Department of Obstetrics & Gynecology, Columbia University Medical Center, New York, NY

Katherine L. Grantz, MD, MS,
Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD

Jagteshwar Grewal, PhD, MPH,
Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD

Sungduk Kim, PhD,
Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD

Robert E. Gore-Langton, PhD,
The Emmes Corporation, Rockville, MD

Germaine M. Buck Louis, PhD, MS,
Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD

Mary E. D'Alton, MD, and
Department of Obstetrics & Gynecology, Columbia University Medical Center, New York, NY

Paul S. Albert, PhD
Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Rockville, MD

Abstract

Objective -—To report on the ultrasound quality assurance (QA) program for the NICHD Fetal Growth Studies and describe both its advantages and generalizability.

Methods -—After training on the Voluson E8 with ViewPoint software (GE Healthcare; Milwaukee, WI), research sonographers were expected to capture blank (unmeasured) images in

Corresponding author: Paul S. Albert, Ph.D., Chief and Senior Investigator, Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health & Human Development, 6100 Executive Blvd., Room 7B05F, Rockville MD 20852, Phone: 301-496-5582, albertp@mail.nih.gov.

triplicate for crown-rump length, biparietal diameter, head circumference, abdominal circumference, and femur length. A primary expert sonographer was designated and validated. A 5% sample (n=740 of 14,785 scans) was randomly selected in three distinct rounds from within strata of maternal body mass index (Round 1 only), gestational age, and research site. Unmeasured images were extracted from selected scans and measured in ViewPoint by the expert sonographer. Correlations and coefficients of variation (CV) were calculated, and the within-measurement standard deviation, i.e., technical error of the measurement (TEM), was calculated.

Results -—The reliability between the site sonographers and the expert was high, with correlations exceeding 0.99 for all dimensions in all rounds. The CV %s showed low variability, with the percentage differences being less than 2%, except for abdominal circumference for Rounds 2 and 3, where it averaged about 3%. Correlations remained high (> 0.90) with increasing fetal size; there was a monotonic increase in TEMs but without a corresponding increase in the CV %.

Conclusions -—Using rigorous procedures for training sonographers, coupled with QA oversight, we determined that the measurements acquired longitudinally for singletons are both accurate and reliable, for establishment of an ultrasound standard for fetal growth.

Keywords

obstetrics; ultrasound; reliability; quality assurance; quality control

INTRODUCTION

The aims of the NICHD Fetal Growth Studies were to establish standards for fetal size-for-gestational age and growth (velocity) in singleton pregnancies for four self-identified racial/ethnic groups in the United States[1] and, in comparison, to describe growth trajectories for the fetuses of healthy obese women and for dichorionic twin fetuses. Critical for establishing an ultrasound standard or a size or velocity reference is the assurance of high-quality measurements and reliability, through a formal quality control (QC) scheme similar to those used in the anthropometry of the living.[2–4] This is especially important for growth velocities, to determine the time needed between measurements for growth to exceed measurement error and in recognizing the timing of growth and peak velocity.[5] Measurements need to be made reliably to minimize intra- and interobserver variation and assure consistency over time.[6]

We defined and implemented two complementary approaches for achieving ultrasound quality in the NICHD Fetal Growth Studies. The first was a QC phase involving rigorous training of site sonographers, standardization of transducer pressure, image acquisition, and caliper placement, and a possibility of re-training if necessary.[7] The second was a unique quality assurance (QA) phase that covered the entire study through re-measurement of randomly selected scans. Initial QC was accomplished before sonographers were credentialed to perform study scans and collect measurement data. Our objective is to report on the ultrasound QA program for the NICHD Fetal Growth Studies and comment on its advantages over conventional interobserver reliability schemes or other methods and on its generalizability to other longitudinal ultrasound studies.

MATERIALS AND METHODS

NICHD Fetal Growth Studies cohorts

Beginning in July 2009 through January 2013, the NICHD Fetal Growth Studies recruited 2,334 low-risk gravidas, with a pregravid body mass index (BMI, kg/m^2) of 19–29.9 kg/m^2, with certain menstrual dates estimated from the date of the last menstrual period (LMP), carrying singletons, and with no pre-existing conditions that could impede fetal growth. The complete list of inclusion and exclusion criteria has been published.[1] Gravidas were recruited from 12 clinical sites in the States of Alabama, California (3 sites), Delaware, Illinois, Massachusetts, New Jersey, New York (2 sites), Rhode Island, and South Carolina and from four self-identified racial/ethnic groups: non-Hispanic white (NHW, n=614); non-Hispanic black (NHB, n=611; Hispanic (n=649); and Asian (n=460). An ultrasound estimate of gestation at recruitment was made and for inclusion in the study had to be between 8–13 completed weeks' gestation and to match the LMP-based gestational age within five days for women between 8–10 completed weeks, six days for those between 11–12 completed weeks, and seven days at 13 weeks.

Once recruited, participants were randomized to four measurement schedules each with up to six study visits: baseline (visit 0) at 8–13 weeks; visit 1 at 16–22 weeks; visit 2 at 24–29 weeks; visit 3 at 30–33 weeks; visit 4 at 34–37 weeks; and visit 5 at 38–41 weeks. A total of 486 obese women (pregravid BMI 30–45 kg/m^2) carrying singletons were also recruited, randomized and measured on the same schedules as the low-risk women. All scans for all participant gravidas were subject to QA, regardless of participant weight or completion status. Approval to enroll subjects was obtained from all clinical sites, and the protocol was also approved by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development human subjects' review board. Participating women gave informed consent before beginning the protocol.

Sonographer credentialing (quality control)

Before data collection was started, the dedicated sonographers at each of the clinical sites attended a multi-day educational program under the direction of the study's central sonology unit (Principal Investigator, Mary E. D'Alton, MD, Columbia University Medical Center, New York, NY) that included didactic and hands-on training in the acquisition of standardized images using study equipment and the performance of standardized measurements according to the procedures outlined in the study manual of operations.[7] All scans were performed on identical equipment (Voluson E8, GE Healthcare, Milwaukee, WI) using a transabdominal curved multi-frequency volume transducer (real-time abdominal (RAB) 4–8 MHz) and endovaginal multi-frequency volume transducer (real-time intracavitary (RIC) 6–12 MHz)

Crown rump length (CRL) was measured along the mid-sagittal long axis as the maximum linear distance between the fetal head and rump. Biparietal diameter (BPD) was measured at the level of the thalami and cavum septa pellucida or the cerebral peduncles as the linear distance from the outer edge of the proximal to the inner edge of the distal skull, while head circumference (HC) was measured at the same level (and often on the same images) using the ellipse function around the outer perimeter of the skull. Abdominal circumference (AC) was measured using the ellipse function circumscribing the actual or projected skin line in the transverse plane at the level of the stomach and the junction of the umbilical vein and portal sinus. Femur length (FL) was measured as the linear distance along the long axis of the femoral diaphysis.

Following training, sonographers were required to submit to the central sonology unit 15 singleton credentialing scans, five in each trimester. Each credentialing scan was reviewed and scored for image quality (magnification, gain, resolution), image plane (including the presence of required landmarks), and caliper placement. Each image was expected to demonstrate the quality, plane, and measurement as specified in the study manual of operations; failure of any of these criteria led to a non-passing grade for the image. Each biometric measure was expected to pass in at least four out of five credentialing scans per trimester, and measures that failed on more than one scan per trimester had to be replaced with a supplemental image of the same biometric measure from the same trimester until the passing criteria were met.

When performing study scans, sonographers were expected to acquire high-quality images with the same resolution, plane and magnification demonstrated in credentialing. Sonographers were also expected to acquire study measurements using proper caliper placement with the linear distance function for straight line measures and the ellipse function for circumferential measures.[7] Measurements were captured in ViewPoint (GE Healthcare) using a fillable mask (i.e., computer screen template) designed specifically for the study to blind the sonographers from previous measurements. Images and the accompanying data were electronically transferred to the image coordinating center (Principal Investigator, Robert E. Gore-Langton, PhD, The Emmes Corporation, Rockville, MD) for storage and further processing.

Quality assurance (QA) image acquisition

The decision was made *a priori* that a randomly-selected 5% of the scans with three replicate images for each measurement would be re-measured for QA. For all study scans at all study sites, sonographers were expected, as often as possible, to acquire and save a “blank” (unmeasured two-dimensional image) before each replicate measurement. This was done only for key dimensions, that is, those most frequently used in estimating gestational age and/or estimations of fetal weight.[8] Unmeasured images were collected for crown-rump length (CRL) and biparietal diameter (BPD) which were measured in triplicate in the first trimester (8–13 completed weeks), corresponding to the baseline study visit (visit 0). BPD, head circumference (HC), abdominal circumference (AC), and femur length (FL) were measured in triplicate the second and third trimesters (visits 1–5, 14 weeks to delivery).

Scan and image management

For implementation of the QA scheme, the image coordinating center created a scan management system specifically for the study QA, so that in re-measurement the expert sonographer (“gold standard”) would be blinded to the characteristics of the participants, particularly the gestational age of the fetus and to the measurements taken by the site sonographers on the same images. As part of the system, a trained curator extracted the blank images from the selected scans and saved them into a separate folder for re-measurement by the expert sonographer. The selected scans were then randomized for order of presentation so that the re-measurement approximated a clinical situation (i.e., not clumped by trimester) and access was limited so that the expert only had access each day to the number of scans that could be expected to be measured in a normal clinical day. The images were measured using ViewPoint (GE Healthcare) on a dedicated workstation or laptop computer unconnected to study ultrasound equipment.

Establishing a gold standard

In preparation for QA, we first established a set of three "gold standard" sonographers (experts), with one designated as primary (GS1) and two back-ups (GS2, GS3). The primary expert was a Registered Diagnostic Medical Sonographer (RDMS) with over 25 years of experience in obstetric ultrasound, and the back-up experts were both Maternal-Fetal Medicine specialists. Measurement reliability was determined among them by re-measurement of the blank images from a randomly-selected set of 30 scans, stratified by study visit, from the on-going study.

Statistical methods

Several different metrics were used in evaluation of reliability both among the experts and in comparison with the site sonographers. For all dimensions, measurements were taken in triplicate (three separate images) and averaged. The average (in mm) and percentage differences (± standard deviation, SD) between the site sonographers and expert were calculated to check for systematic measurement differences. Using linear mixed models, the intraclass correlation coefficient (ICC) to validate the three experts and the coefficient of variation (CV) were calculated. The CV, or relative SD, was then converted to a percentage (CV %). The comparison between the site sonographers and expert was made using a Pearson's correlation (*r*).

As an indicator of measurement accuracy, the intraobserver technical error of measurement (TEM) was calculated.[3,4] The TEM, also known as the measurement error standard deviation or unreliability SD, for triplicate measures was calculated using a linear mixed model with a single random intercept where the TEM is the residual standard error in the model. The TEM can be interpreted much like a SD where approximately two-thirds of measurements should be within the mean ± TEM for each triplicate set.

RESULTS

Establishing a gold standard

The initial validation of the primary gold standard (GS1) sonographer was accomplished before implementation of QA by comparison of GS1 with two secondary experts (GS2 and GS3), who were then available as back-ups. The blank (unmeasured) images from 30 scans, randomly selected to represent all visits, were measured independently by the three experts (Table 1). Accuracy among the three sonographers was excellent, with ICCs exceeding 0.99 for every dimension and CV %s (CV × 100) showing low variation. The CV %s were less than 2% for HC and FL and less than 1% for CRL, BPD, and AC. The TEMs were comparable across the experts.

Implementation of the QA for singletons

For ease of management, interim review, and remediation if necessary, there were three rounds of QA assessment. The total number of scans available for QA by the end of the study was 14,785, representing scans from 2,820 individual gravidas (Table 2). Exactly 5.0% of the scans (*n*=740) were selected for QA review, representing 26.2% of the study participants with each participant represented only once for purposes of QA. Scans in each round were selected for QA based on cut-offs for expected date of delivery, so that a gravida's complete set of scans was available for selection. Scans from participants who had not delivered or were not expected to deliver by the cut-off dates were included in the subsequent round. The cut-off for Round 1 was expected delivery by October 1, 2011; for Round 2, expected delivery by October 1, 2012; for Round 3, study completion (October 1, 2013). The scans in Round 1 were randomized and selected based on strata for cohort (low-risk, obese), visit number (0, 1, 2, 3, 4, 5), and sonographer/site, while Rounds 2 and 3 were stratified by only visit number and sonographer/site because there were too few in the obese cohort for a separate stratum and because analysis of Round 1 indicated that there was no difference in measurement reliability by maternal weight status (data not shown).

There were fewer blank images available for QA than there were replicate measurements, although the number of missing blank images was not large (1.5%). Of the 7,707 expected images (3 images per measure for each scan), there were only 16 missing measured images (0.2%) for the site sonographers, but 115 missing unmeasured images (1.5%).

The reliability between the site sonographers and GS1 remained high over the course of the study, with the overall correlations exceeding 0.99 for all dimensions in all Rounds (Table 3). The concordance for CRL from 8–13 completed weeks ($r = 0.994$, $P < .001$) is shown in Figure 1. The concordance for BPD from 8–41 completed weeks ($r = 0.999$, $P < .001$) and for FL ($r = 0.998$, $P < .001$), HC ($r = 0.998$, $P < .001$), and AC ($r = 0.996$, $P < .001$) from 14–41 completed weeks are shown in Figure 2. The CV %s for all measures showed low variability, and the percentage differences were less than 2%, with the exception of the measurement of AC for Round 2 and Round 3, where the percentage difference averaged about 3%. However, for both the site sonographers and for GS1 the TEM remained consistent over the Rounds, despite there being more scans with fewer than three images available for re-measurement.

AC measurement in Rounds 2 and 3 were less reliable than the other key dimensions, with GS1 measuring larger than the site sonographers. However, the difference was consistent for all sites, sonographers, and fetal size, indicating that GS1 was the source of the difference, not the site sonographers.

Likewise, correlations remained high (> 0.90, $P < .001$) between the site sonographers and GS1 through gestation with increasing fetal size and without a significant increase in the CV % (Table 4). For most measurements there was a trend toward a greater absolute discrepancy with advancing gestational age, but this did not translate into an increased percentage difference or higher CV %s. There were also no differences by site sonographers, such that all site sonographers measured equally well, relative to the primary expert.

To assure the on-going accuracy of the primary gold standard (GS1), after each round a random set of 30 scans from the Round were measured by one of the secondary gold standards (GS2) with the finding that the primary gold standard (GS1) remained reliable over time, although the measurement of AC drifted slightly higher by about 5–6 mm (data not shown).

DISCUSSION

Quality control (QC) v. quality assurance (QA)

Given the nature of image acquisition and software, there are two approaches to measurement reliability that can be applied in ultrasound studies. Both presuppose sonographers' rigorous training in image acquisition and caliper placement before data collection begins, similar to the training needed to credential anthropometrists in positioning subjects and locating landmarks. We opted for a QA scheme that was on-going throughout the study and was developed because the benefits of on-going QA outweighed those of implementing a QC plan of on-site interobserver reliability[9,10] or of monitoring the measurement variability (intraobserver reliability) of site sonographers.[11–13] The latter presumes that the TEMs and tolerable variability of the biometric measures are established so that the intraobserver reliability can be monitored prospectively using cumulative sum (CUSUM) graphs or other graphical techniques,[12,13] and is implemented more for QC than for QA.

We also sought to minimize inconvenience to participants and to limit the amount of ultrasound exposure, while at the same time blinding the sonographers to which scans would be selected, equally exposing all scans to selection for QA, and blinding the expert ("gold standard") sonographer to participant characteristics that could affect measurement (e.g., race/ethnicity, maternal weight status, gestational age of the fetus).[14,15]

QC by periodic interobserver measurement of the same participant[10,16,17] is by far the most common in ultrasound studies, and there are benefits to this approach. Issues can be identified and sonographers re-trained immediately with assurance going forward. There are, however, a number of drawbacks. There is a need for experts to oversee data collection remotely or travel in multisite studies and/or to establish reliability among local experts. The acquisition of multiple scans is time-consuming for participants, especially in the third trimester, at least doubling the length of the visit. For research purposes, it may be necessary to obtain additional consent for a reliability scan, and the sample re-scanned may not be representative (i.e., representing all factors that may affect measurement, such as gestational age) or present when the expert is available on-site or remotely. Finally, re-scanning the same gravida may introduce artificially high agreement by having sonographers aware that participant will be re-scanned. That is, the sonographer is not blinded to which participants are selected for reliability.[17] A novel scheme using archived three-dimensional volumes to re-measure for two-dimensional reliability was recently employed by the INTERGROWTH-21st study, and while this may overcome some of the issues associated with conventional QC, it was noted that acquisition of the volumes added considerable time to the visit, and there was some unanticipated difficulty in manipulating the three-dimensional volumes and locating the correct planes or positions for measuring some of the linear dimensions.[18] For example, measurements were systematically smaller using three-dimensional volumes acquired when the head was facing anterior and posterior or when the long axis of the femur was perpendicular to the transducer.

The NICHD Fetal Growth Studies instead had a planned *post hoc* determination of reliability, i.e., QA accomplished by measurement of blank (unmeasured) images from randomly-selected scans. In this case, the emphasis was more on caliper placement and measurement than image acquisition. However, quality image acquisition was a core component of sonographer training, using a scheme of image scoring similar to that developed for the First And Second Trimester Evaluation of Risk (FASTER) Trial that evaluated the quality of nuchal translucency measures.[19] It was planned for the NICHD Fetal Growth Studies that insofar as possible all sonographers were to save blanks (unmeasured images) before each measurement for key dimensions for experts to re-measure. This has a number of benefits over conventional QC. First, expert sonographers can work from a central location or electronically using the saved images and can do so expeditiously, so that many more QA images covering more contingencies (e.g., advancing gestational age, race/ethnicity, maternal weight, sonographer, clinical site) can be assessed. The acquisition of blank images may add some time to the study visit, but it is minimal and all patients are equally exposed with no additional consent needed.

There are other advantages. Every scan is equally subject to QA (within limits of stratification), and the sample for reliability can be structured to ensure that factors thought to affect scan acquisition and measurement can be equally represented, e.g., maternal weight group, gestational age,[14,15] and sonographer. Importantly, if blanks are available, it is possible to score scan quality independently and re-measure particular dimensions or exclude individual questionable measures, such as an AC in the third trimester, if it is found that the scans and/or measurements were problematic for certain sonographers. We were also able to establish TEMs by gestational age for real-time evaluation of measurement variation in future studies. Finally, this scheme ensures that the sonographer is blinded to which scans and participants will be re-measured for reliability, and the experts are blinded to the sonographers' measurements.

Limitations

There are some limitations to the QA scheme that was developed. *Post hoc*, there was no real-time oversight of sonographers while performing research scans, meaning that there is more reliance on sonographers' initial training for image acquisition. This reinforces the importance of credentialing sonographers when beginning ultrasound studies.[7] There may be reasons, such as visit length or fetal position, that will make the acquisition of unmeasured images difficult, and that will affect the estimation of reliability. A QA scheme that is primarily assessing caliper placement does not address the quality of image acquisition (i.e., image quality, landmark presence, magnification), although those can be assessed qualitatively on the blanks in addition to the quantitative assessment. We were unable to detect in a timely fashion a minor change in the measurement technique of the primary expert, so that a recommendation from this study would be to have more frequent standardization among experts. Finally, remediation by re-measuring saved images could be time-consuming and costly, but at least it is feasible. Still, in the context of our study with good training of sonographers, there was no need for remediation of any site sonographer and the QA approach worked well.

Summary of QA findings

Using rigorous procedures for training and credentialing sonographers, coupled with QA oversight of a 5% random selection of scans, we determined that the measurements acquired longitudinally for singletons in the NICHD Fetal Growth Studies are both accurate and reliable, minimizing measurement and sonographer error, for establishment of a standard for fetal growth. Specifically, the low measurement variability and technical errors of the measurement reinforce the validity of the trajectories and significance of the racial/ethnic differences in fetal growth that we observed.[1] We also found that, of the measurements used most commonly in equations to estimate fetal weight, AC (a soft tissue measure) is the least reliable and most variable, which should be taken into account in models and studies that emphasize AC or AC velocity as a major predictor of fetal outcome.[20]

There are unique features to QA by re-measurement of saved images that make it well worth consideration for future studies. First, sonographers are blinded to what scans will be used for QA, and the experts are blinded to characteristics that may affect measurement (e.g., maternal weight, race/ethnicity). Second, if it is found that a particular sonographer is measuring unreliably, it is possible to re-measure archived scans to minimize error. Finally, reliable measurements have value clinically in accurately sizing fetuses and defining abnormal growth to ensure proper diagnosis and timely intervention.

Acknowledgments

We would like to thank the Principal Investigators at the clinical sites for their support and the site sonographers for their good and careful work. The Principal Investigators are: Anthony Sciscione, DO (Christiana Care Health System, Delaware); Deborah A. Wing, MD (University of California, Irvine, Long Beach Memorial Medical Center, and Fountain Valley Regional Hospital and Medical Center); William A. Grobman, MD, MBA (Northwestern University Feinberg School of Medicine, Chicago, Illinois); Roger B. Newman, MD (Medical University of South Carolina); Ronald Wapner, MD (Columbia University Medical Center, New York City); Daniel Skupski, MD (New York Hospital Queens); Angela C. Ranzini, MD (St. Peter's University Hospital, New Brunswick, New Jersey); John Owen, MD, MSPH (University of Alabama at Birmingham); Edward K. Chien, MD (Women and Infants Hospital of Rhode Island); and Sabrina Craigo, MD (Tufts Medical Center, Boston, Massachusetts).

We would like to especially acknowledge the contributions of Patricia Morgan, RDMS, who served as the primary expert sonographer, and George Bega, MD, who served as a back-up expert for QA. Matthew Beyers (The Emmes Corporation) was instrumental in setting up the image management system to conduct the QA, and Scott McCrimmon (The Emmes Corporation) contributed his invaluable technical expertise in image management. This research was supported by funding through the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development intramural program, contracts HHSN275200800013C, HHSN2752008000002I, HHSN27500006, HHSN275200800003IC, HHSN275200800014C, HHSN275200800012C, HHSN275200800028C, HHSN275201000009C, and included funding through the American Recovery and Reinvestment Act of 2009.

REFERENCES

1. Buck Louis GM, Grewal J, Albert PS, Sciscione A, Wing DA, Grobman WA, Newman RB, Wapner R, D'Alton ME, Skupski D, Nageotte MP, Ranzini AC, Owen J, Chien EK, Craigo S, Hediger ML, Kim S, Zhang C, Grantz KL. Racial/ethnic standards for fetal growth, the NICHD Fetal Growth Studies. Am J Obstet Gynecol 2015; 213(4):449.e1–449.e41. [PubMed: 26410205]
2. Mueller RH, Martorell R. Reliability and accuracy of measurement In: Lohman TG, Roche AF, Martorell R (eds), Anthropometric Standardization Reference Manual. Champaign, IL: Human Kinetics Books, 1988:83–86.
3. Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. Br J Nutr 1999; 82:165–177 (Erratum: *Br J Nutr* 2000; 83:95). [PubMed: 10655963]
4. Cameron N Essential anthropometry: baseline anthropometric methods for human biologists in laboratory and field situations. Am J Hum Biol 2013; 25:291–299. [PubMed: 23606226]
5. Milani S, Bossi A, Bertino E, Di Battista E, Coscia A, Aicardi G, Fabris C, Benso L. Differences in size at birth are determined by differences in growth velocity during early prenatal life. Pediatr Res 2005; 57:205–210. [PubMed: 15611356]
6. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. Ultrasound Obstet Gynecol 2008; 31:466–475. [PubMed: 18306169]
7. Fuchs KM, D'Alton M, for the NICHD Fetal Growth Study. Can sonographer education and image review standardize image acquisition and caliper placement in 2D ultrasounds? Experience from the NICHD Fetal Growth Study. Am J Obstet Gynecol 2012; 206(1 suppl):S15–S16.
8. Anderson NG, Jolly IJ, Wells JE. Sonographic estimation of fetal weight: comparison of bias, precision and consistency using 12 different formulae. Ultrasound Obstet Gynecol 2007; 30:173–179. [PubMed: 17557378]
9. Solomon LJ, Bernard JP, Ville Y. Analysis of Z-score distribution for the quality control of fetal ultrasound measurements at 20–24 weeks. Ultrasound Obstet Gynecol 2005; 26:750–754. [PubMed: 16308899]
10. Lima JC, Miyague AH, Filho FM, Nastri CO, Martins WP. Biometry and fetal weight estimation by two-dimensional and three-dimensional ultrasonography: an intraobserver and interobserver reliability and agreement study. Ultrasound Obstet Gynecol 2012; 40:186–193. [PubMed: 22102507]
11. 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; International Fetal and Newborn Growth Consortium for the 21st Century. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. Ultrasound Obstet Gynecol 2011; 38:681–687. [PubMed: 22411446]
12. Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, Altman DG, Papageorghiou AT; International Fetal and Newborn Growth Consortium for the 21st Century (INTERGROWTH-21st). Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet Gynecol 2012; 39:266–273. [PubMed: 22535628]
13. Sarris I, Ioannou C, Ohuma EO, Altman DG, Hoch L, Cosgrove C, Fathima S, Salomon LJ, Papageorghiou AT; International Fetal and Newborn Growth Consortium for the 21st Century. Standardisation and quality control of ultrasound measurements taken in the INTERGROWTH-21st Project. BJOG 2013; 120(suppl 2):33–37. [PubMed: 23841486]
14. Deter RL, Harrist RB, Hadlock FP, Carpenter RJ. Fetal head and abdominal circumferences. II. A critical re-evaluation of the relationship to menstrual age. J Clin Ultrasound 1982; 10:365–372. [PubMed: 6816816]
15. Harstad TW, Buschang PH, Little BB, Santos-Ramos R, Twickler D, Brown CEL. Ultrasound anthropometric reliability. J Clin Ultrasound 1994; 22:531–534. [PubMed: 7806660]
16. Perni SC, Chervenak FA, Kalish RB, Magherini-Rothe S, Preganic M, Streltzoff J, Skupski DW. Intraobserver and interobserver reproducibility of fetal biometry. Ultrasound Obstet Gynecol 2004; 24:654–658. [PubMed: 15476300]

17. Coelho Neto MA, Roncato P, Nastri CO, Martins WP. True Reproducibility of UltraSound Techniques (TRUST): systematic review of reliability studies in obstetrics and gynecology. Ultrasound Obstet Gyncol 2015; 46:14–20.

18. Sarris I, Ohuma E, Ioannou C, Sande J, Altman DG, Papageorghiou AT, International Fetal and Newborn Growth Consortium for the 21st Century. Fetal biometry: how well can offline measurements from three-dimensional volumes substitute real-time two-dimensional measurements? Ultrasound Obstet Gynecol 2013; 42:560–570. [PubMed: 23335102]

19. D'Alton ME, Cleary-Goldman J, Lambert-Messerlian G, Ball RH, Nyberg DA, Comstock CH, Bukowski 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. [PubMed: 19173241]

20. 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 Published online 9 7, 2015. DOI: 10.1016/S0140-6736(15)00131-2.

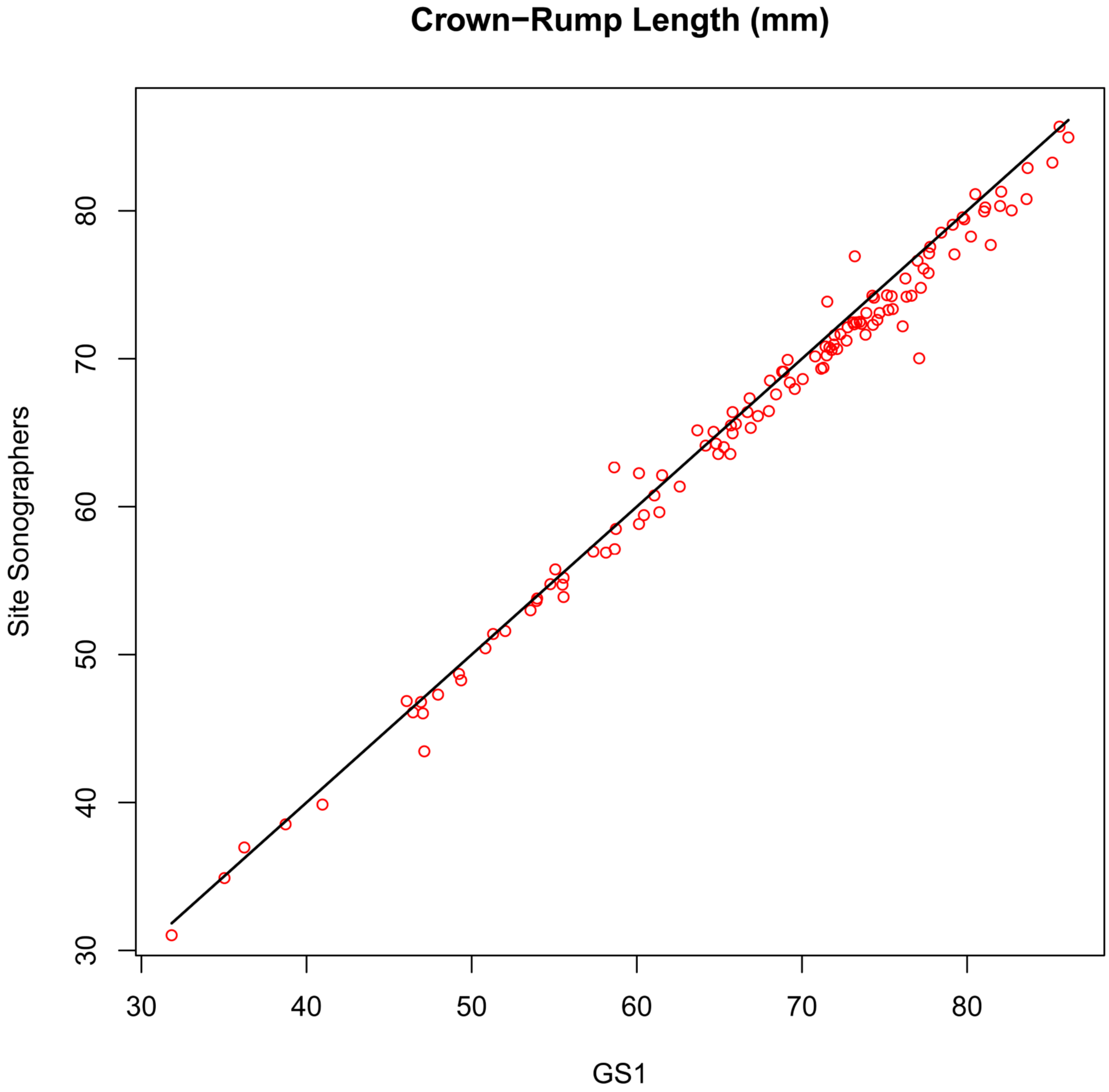

Figure 1.

Measurement concordance (*r* = 0.994, *P* < .001) for crown-rump length (CRL) between the primary expert sonographer (GS1, *x*-axis) and the site sonographers (*y*-axis).

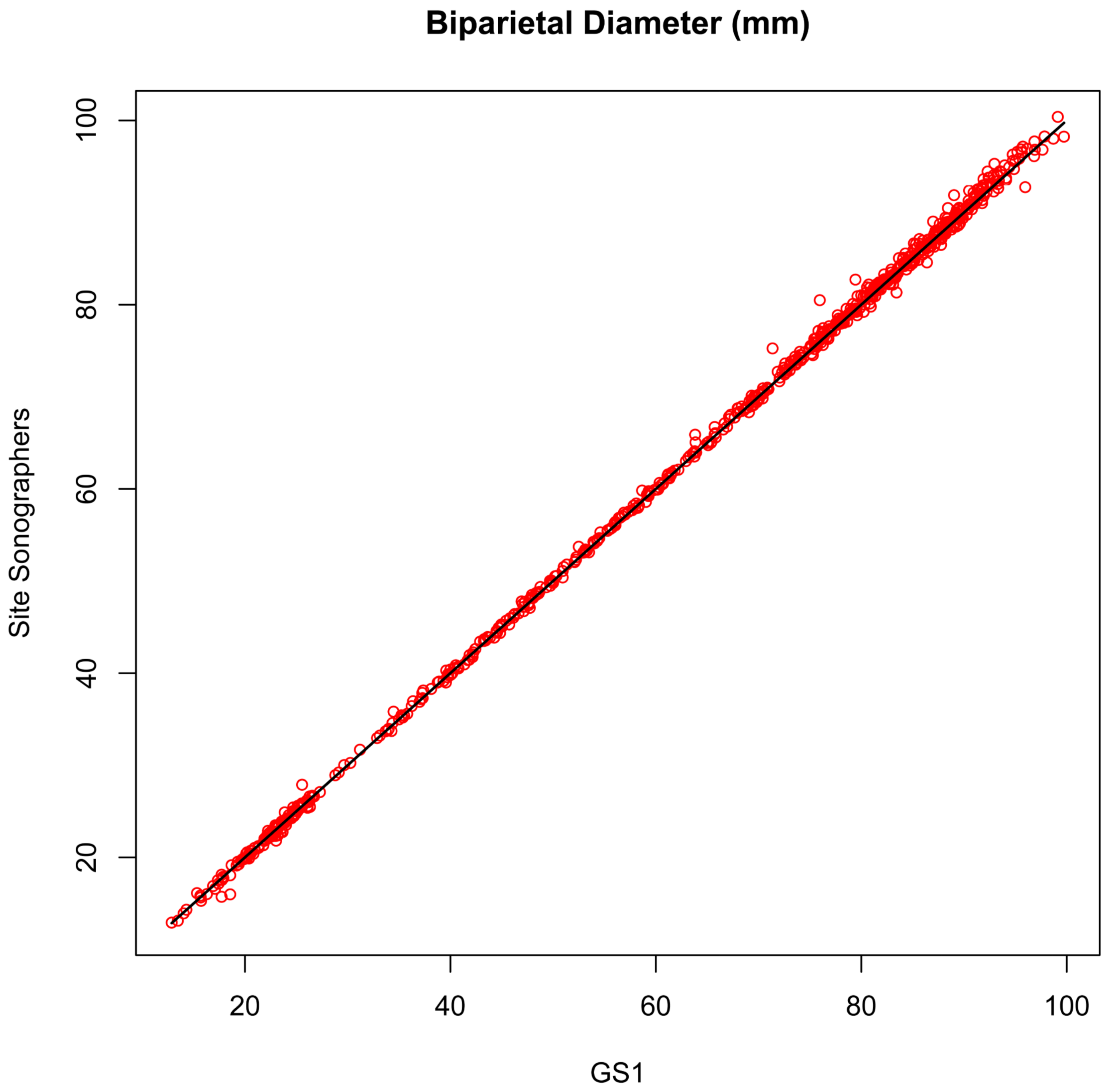

Figure 2a.

Measurement concordance ($r = 0.999$, $P < .001$) for biparietal diameter (BPD, outer-inner) between the primary expert sonographer (GS1, *x*-axis) and the site sonographers (*y*-axis).

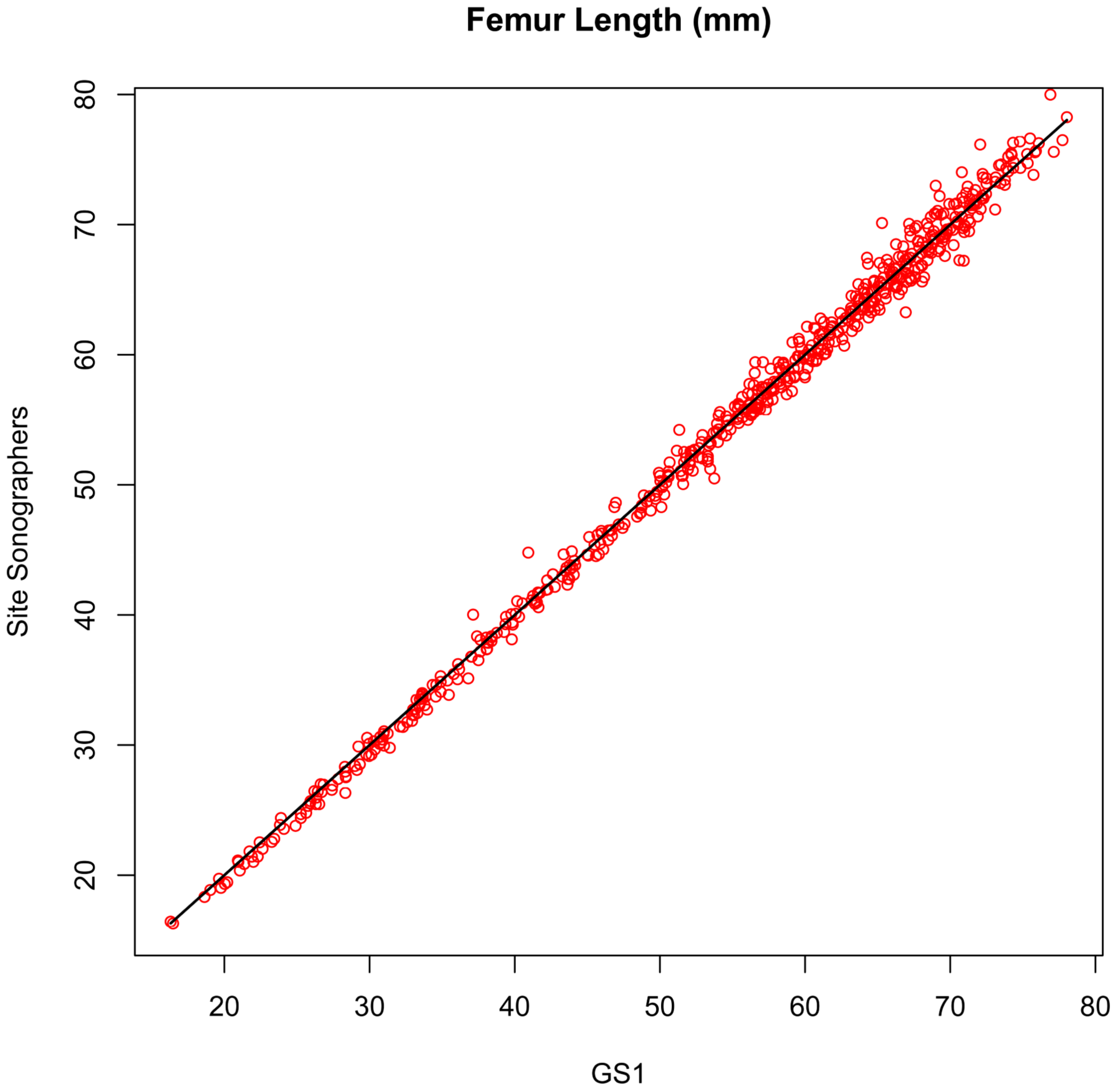

Figure 2b.

Measurement concordance (*r* = 0.998, *P* < .001) for femur length (FL) between the primary expert sonographer (GS1, *x*-axis) and the site sonographers (*y*-axis).

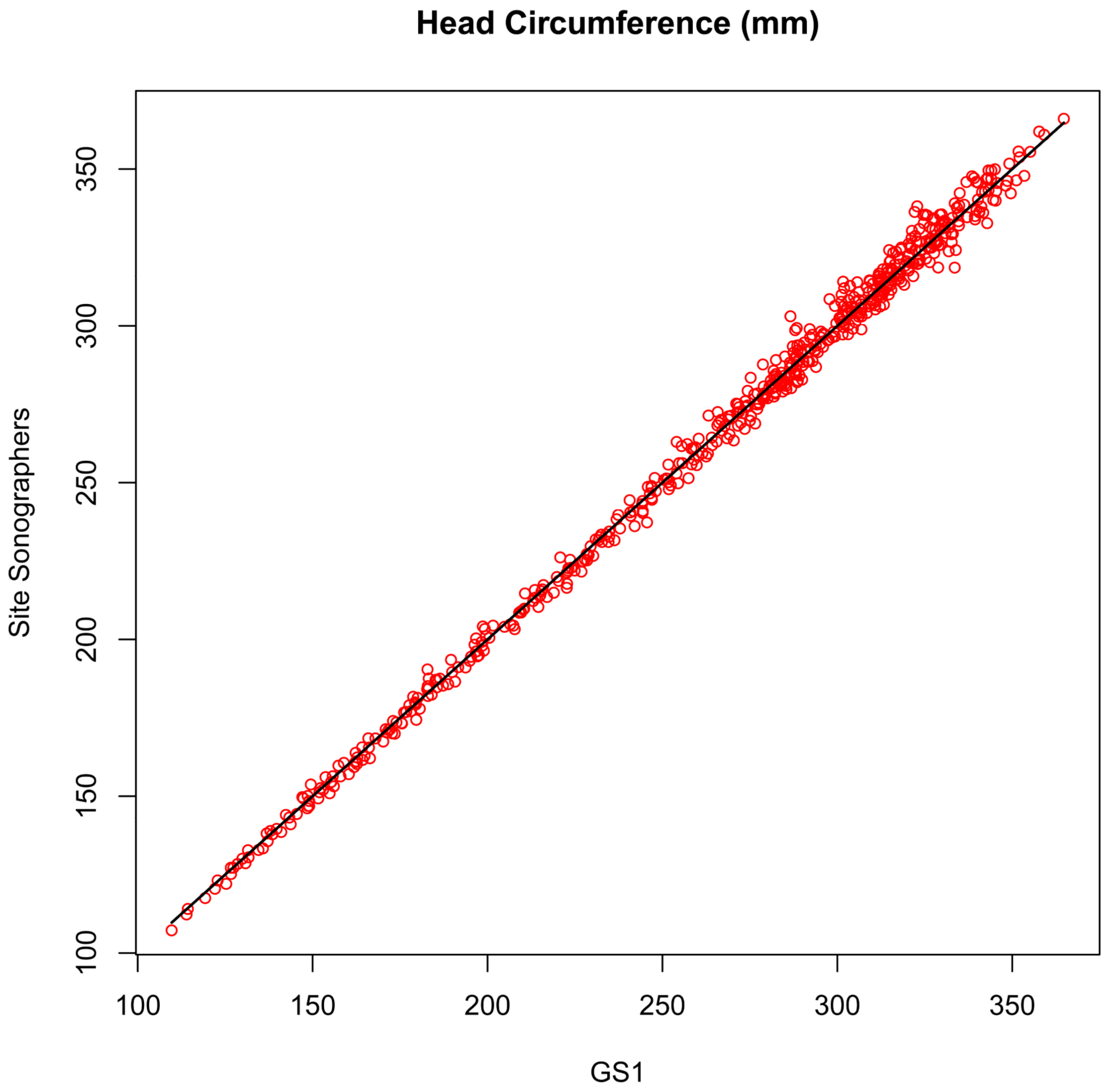

Figure 2c.

Measurement concordance ($r = 0.998$, $P < .001$) for head circumference (HC) between the primary expert sonographer (GS1, x-axis) and the site sonographers (y-axis).

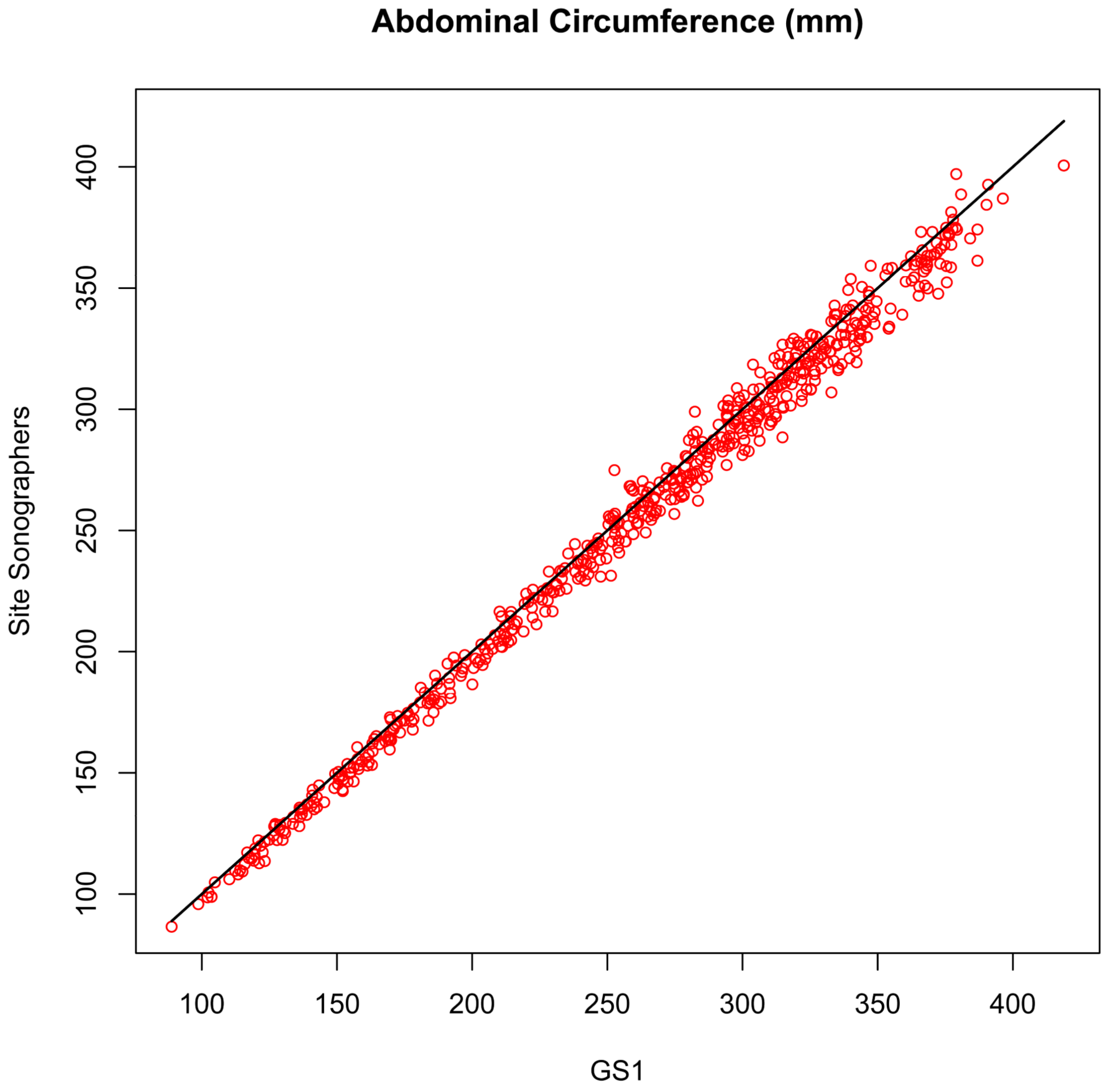

Figure 2d.

Measurement concordance ($r = 0.996$, $P < .001$) for abdominal circumference (AC) between the primary expert sonographer (GS1, *x*-axis) and the site sonographers (*y*-axis).

Table 1.

Initial Validation of Primary Gold Standard (GS1) Sonographer with Secondary Gold Standard Sonographers (GS2, GS3)

Dimension	**n**	**GS1 mean ± SD**	**Difference from GS2 (mm) mean ± SD**	**Difference from GS3 (mm) mean ± SD**	**Difference from GS2 (%) mean ± SD**	**Difference from GS3 (%) mean ± SD**	**GS1 TEM**	**GS2 TEM**	**GS3 TEM**	**ICC**	**CV %**
CRL (mm)	6	68.9 ± 9.5	−0.1 ± 0.3	−0.7 ± 0.5	−0.03 ± 0.4	−1.0 ± 0.7	1.3	1.4	1.3	0.99	0.7
BPD (mm)	30	63.6 ± 27.1	0.2 ± 0.3	0.8 ± 0.6	0.2 ± 0.6	1.4 ± 1.0	0.6	0.5	0.7	0.99	0.9
HC (mm)	24	268.1 ± 72.1	0.9 ± 2.1	−2.6 ± 4.5	0.4 ± 0.9	−1.2 ± 1.7	1.7	1.7	1.8	0.99	1.3
AC (mm)	24	262.7 ± 85.1	−3.8 ± 3.1	−2.6 ± 7.3	−0.9 ± 1.6	−1.2 ± 2.5	1.9	2.7	2.2	0.99	0.8
FL (mm)	24	53.4 ± 17.1	−0.5 ± 1.2	0.1 ±0.8	−1.0 ± 2.0	0.3 ± 1.7	0.7	0.6	0.8	0.99	1.6

The differences were calculated as (GS2 – GS1) and (GS3 – GS1) such that a positive value indicates that the GS2 or GS3 measurements were larger and a negative value that they were smaller. Linear mixed models were used to estimate the intraclass correlation coefficient (ICC) and the coefficient of variation (CV) among the three sonographers. AC, abdominal circumference; BPD, biparietal diameter; CRL, crown-rump length; CV %, coefficient of variation percent (CV × 100); FL, femur length; GS, gold standard; HC, head circumference; ICC, intraclass correlation coefficient; and TEM, technical error of measurement.

Table 2.

Comparison of the Characteristics of the Total Number of Scans Acquired for Singletons with Those Randomly Selected for Quality Assurance (QA)

Characteristic	***Specifics***	***Total sample***	***QA sample***
		n	**n (%)**
Scans acquired			
	Total all scans	14,785	740 (5.0)
	Individual gravidas	2820	740 (26.2)
QA rounds			
1	Delivered by 01 Oct 2011	6552	328 (5.0)
2	Delivered by 01 Oct 2012	5671	284 (5.0)
3	Delivered by 01 Oct 2013	2562	128 (5.0)
Cohorts			
	Low-risk (BMI 19-29.9 kg/m^2)	12,356	576 (4.7)
	Obese (BMI ≥ 30 kg/m^2)	2429	164 (6.8)
Geographical regions			
Northeast		4391	224 (5.1)
South		5690	285 (5.0)
Midwest		1860	92 (4.9)
West		2844	139 (4.9)
Study visit	Gestation		
0	8-13 weeks	2799	134 (4.8)
1	16-22 weeks	2710	141 (5.2)
2	24-29 weeks	2600	128 (5.0)
3	30-33 weeks	2545	132 (5.2)
4	34-37 weeks	2429	120 (4.9)
5	38-41 weeks	1702	84 (4.9)

Data for the total number of scans are the number for the total number of scans acquired, not the total number of individual gravidas. Data for the QA sample are given as n (%), representing the number and percentage of the total number of scans. Individual gravidas (n=2820) could be represented up to 6 times in the total number of scans, but only once in the QA. BMI, pregravid body mass index (kg/m^2); and QA, quality assurance.

Table 3.

Ultrasound Measurement Reliability by Comparison of Site Sonographers (SS) with Primary Gold Standard (GS1)

Dimension	n	SS mean ± SD	GS1 mean ± SD	Difference from GS1 (mm) mean ± SD	Difference from GS1 (%) mean ± SD	SS TEM	GS1 TEM	Correlation (*r*)	CV %
				Round 1					
CRL (mm)	46	67.9 ± 13.0	68.3 ± 13.3	0.4 ± 1.2	0.6 ± 1.8	1.7	1.6	0.99	1.3
BPD (mm)	295	65.1 ± 23.9	65.0 ± 23.9	−0.2 ± 0.6	−0.2 ± 1.0	0.8	0.7	0.99	0.7
HC (mm)	253	267.3 ± 61.0	267.8 ± 61.0	0.6 ± 3.8	0.2 ± 1.4	3.2	2.2	0.99	1.0
AC (mm)	252	253.5 ± 72.3	254.5 ± 71.8	0.9 ± 5.8	0.5 ± 2.2	4.9	3.7	0.99	1.6
FL (mm)	250	54.1 ± 15.2	53.9 ± 14.9	−0.2 ± 1.0	−0.1 ± 1.9	0.9	0.8	0.99	1.3
				Round 2					
CRL (mm)	43	64.9 ±10.8	65.9 ± 11.3	1.1 ± 1.4	1.6 ± 2.1	1.5	1.4	0.99	1.9
BPD (mm)	257	62.0 ± 25.4	61.8 ± 25.3	−0.2 ± 0.7	−0.2 ± 1.6	0.7	0.6	0.99	0.8
HC (mm)	206	263.7 ± 63.9	262.8 ± 62.8	−0.9 ± 4.0	−0.3 ± 1.4	3.0	1.9	0.99	1.1
AC (mm)	210	254.9 ± 79.4	261.7 ± 80.5	6.8 ± 5.9	2.7 ± 2.0	5.0	2.9	0.99	2.5
FL (mm)	211	54.0 ± 15.6	54.2 ± 15.5	0.3 ± 1.0	0.6 ± 1.9	1.0	0.9	0.99	1.4
				Round 3					
CRL (mm)	19	65.7 ± 10.2	66.8 ± 10.3	1.1 ± 0.8	1.7 ± 1.2	1.5	1.6	0.99	1.4
BPD (mm)	113	64.6 ± 25.0	64.3 ± 24.8	−0.3 ± 0.5	−1.3 ± 0.9	0.6	0.6	0.99	0.7
HC (mm)	99	271.8 ± 58.8	271.8 ± 58.0	−0.1 ± 3.4	−0.05 ± 1.2	2.9	2.1	0.99	0.9
AC (mm)	98	260.0 ± 71.2	269.4 ± 73.3	9.5 ± 6.3	3.6 ± 2.1	5.2	3.0	0.99	3.0
FL (mm)	96	55.0 ± 14.8	55.3 ± 14.4	0.2 ± 1.1	0.6 ± 1.8	1.1	1.0	0.99	1.4

The differences were calculated as (GS1 – Site Sonographer) such that a positive value indicates that the Site Sonographers’ measurements were smaller and a negative value that they were larger. AC, abdominal circumference; BPD, biparietal diameter; CRL, crown-rump length; CV %, coefficient of variation percent (CV × 100); FL, femur length; GS1, gold standard sonographer 1; HC, head circumference; SS, site sonographers; and TEM, technical error of measurement.

Table 4.

Ultrasound Measurement Reliability by Gestation Comparing the Site Sonographers (SS) with Primary Gold Standard (GS1) Sonographer

Dimension	***Gestation* (wk)**	**n**	**SS mean ± SD**	**Difference from GS1 (mm) mean ± SD**	**Difference from GS1 (%) mean ± SD**	**SS TEM**	**GS1 TEM**	**Correlation (*r*)**	**CV %**
CRL (mm)	8-13	108	66.3 ± 11.7	0.8 ± 1.3	1.2 ± 1.9	1.6	1.5	0.99	1.6
BPD (mm)	8-13	113	22.0 ± 3.4	0.04 ± 0.5	0.2 ± 2.4	0.6	0.4	0.99	1.6
	16-22	128	46.2 ± 8.2	−0.1 ± 0.3	−0.3 ± 0.7	0.5	0.4	0.99	0.5
	24-29	117	67.5 ± 7.0	−0.3 ± 0.6	−0.4 ± 0.9	0.6	0.7	0.99	0.7
	30-33	125	79.9 ± 4.2	−0.2 ± 0.6	−0.3 ± 0.8	0.7	0.6	0.99	0.6
	34-37	113	87.3 ± 4.1	−0.3 ± 0.7	−0.3 ± 0.8	0.9	0.8	0.99	0.6
	38-41	69	91.2 ± 4.2	−0.3 ± 0.9	−0.4 ± 1.0	1.0	1.0	0.98	0.7
HC (mm)	16-22	128	172.0 ± 29.7	0.5 ± 2.2	0.3 ± 1.3	1.8	1.4	0.99	0.9
	24-29	119	250.8 ± 23.1	0.6 ± 3.1	0.2 ± 1.2	2.7	1.5	0.99	0.9
	30-33	125	293.6 ± 13.9	0.04 ± 4.2	0.0 ± 1.4	3.2	2.1	0.95	1.0
	34-37	115	319.4 ± 13.6	−1.0 ± 4.6	−0.3 ± 1.5	3.7	2.6	0.94	1.0
	38-41	71	331.6 ± 13.9	−1.4 ± 6.4	−0.5 ± 2.0	4.0	2.5	0.90	1.3
AC (mm)	16-22	133	149.2 ± 28.8	3.6 ± 3.5	2.4 ± 2.2	3.4	2.4	0.99	2.3
	24-29	119	228.0 ± 25.6	4.6 ± 5.2	2.0 ± 2.2	3.7	3.0	0.98	2.1
	30-33	121	279.6 ± 21.3	5.0 ± 7.1	1.7 ± 2.5	5.3	3.5	0.95	2.2
	34-37	111	320.4 ± 22.0	5.0 ± 8.7	1.5 ± 2.7	6.3	4.1	0.93	2.2
	38-41	76	349.0 ± 22.3	5.3 ± 9.5	1.5 ± 2.7	6.3	3.4	0.91	2.2
FL (mm)	16-22	129	31.3 ± 7.3	0.4 ± 0.5	1.3 ± 1.8	0.7	0.8	0.99	1.5
	24-29	118	50.1 ± 5.3	0.1 ± 1.0	0.1 ± 2.0	0.9	0.9	0.98	1.4
	30-33	122	59.9 ± 4.1	0.1 ± 1.0	0.1 ± 1.6	0.9	0.8	0.97	1.2
	34-37	110	67.1 ± 4.0	−0.1 ±1.2	−0.9 ± 1.8	1.0	1.0	0.95	1.3
	38-41	78	71.2 ± 3.5	−0.3 ± 1.4	−0.5 ± 2.0	1.2	1.2	0.92	1.4

The differences were calculated as (GS1 – Site Sonographer) such that a positive value indicates that the Site Sonographers' measurements were smaller and a negative value that they were larger. AC, abdominal circumference; BPD, biparietal diameter; CRL, crown-rump length; CV %, coefficient of variation percent (CV × 100); FL, femur length; GS1, gold standard sonographer 1; HC, head circumference; SS, site sonographers; and TEM, technical error of measurement.