2012

Biometry and fetal weight estimation by two-dimensional and three-dimensional ultrasonography: an intraobserver and interobserver reliability and agreement study

ULTRASOUND IN OBSTETRICS & GYNECOLOGY, HOBOKEN, v. 40, n. 2, pp. 186-193, AUG, 2012
http://www.producao.usp.br/handle/BDPI/42065

Ultrasound Obstet Gynecol 2012; **40**: 186–193
Published online 3 July 2012 in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.10146

Biometry and fetal weight estimation by two-dimensional and three-dimensional ultrasonography: an intraobserver and interobserver reliability and agreement study

J. C. LIMA*†, A. H. MIYAGUE*†, F. M. FILHO*†, C. O. NASTRI*† and W. P. MARTINS*†‡

**Departamento de Ginecologia e Obstetrícia da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (FMRP-USP), Ribeirão Preto, Brazil; †Escola de Ultra-sonografia e Reciclagem Médica de Ribeirão Preto (EURP), Ribeirão Preto, Brazil; ‡Instituto Nacional de Ciência e Tecnologia (INCT) de Hormônios e Saúde da Mulher, Ribeirão Preto, Brazil*

KEYWORDS: fetal weight; reproducibility of results; ultrasonography

ABSTRACT

Objective *To evaluate and compare the intraobserver and interobserver reliability and agreement for the biparietal diameter (BPD), abdominal circumference (AC), femur length (FL) and estimated fetal weight (EFW) obtained by two-dimensional ultrasound (2D-US) and three-dimensional ultrasound (3D-US).*

Methods *Singleton pregnant women between 24 and 40 weeks were invited to participate in this study. They were examined using 2D-US in a blinded manner, twice by one observer, intercalated by a scan by a second observer, to determine BPD, AC and FL. In each of the three examinations, three 3D-US datasets (head, abdomen and thigh) were acquired for measurements of the same parameters. We determined EFW using Hadlock's formula. Systematic errors between 3D-US and 2D-US were examined using the paired* t*-test. Reliability and agreement were assessed by intraclass correlation coefficients (ICCs), limits of agreement (LoA), SD of differences and proportion of differences below arbitrary points.*

Results *We evaluated 102 singleton pregnancies. No significant systematic error between 2D-US and 3D-US was observed. The ICC values were higher for 3D-US in both intra- and interobserver evaluations; however, only for FL was there no overlap in the 95% CI. The LoA values were wider for 2D-US, suggesting that random errors were smaller when using 3D-US. Additionally, we observed that the SD values determined from 3D-US differences were smaller than those obtained for 2D-US. Higher proportions of differences were below the arbitrarily defined cut-off points when using 3D-US.*

Conclusion *3D-US improved the reliability and agreement of fetal measurements and EFW compared with 2D-US.*

INTRODUCTION

Estimation of fetal weight is routinely performed in obstetric practice and is used to assist in the decision-making process regarding the timing of delivery of growth-restricted and macrosomic fetuses[1,2]. Several mathematical formulae can be used to calculate estimated fetal weight (EFW). Among them, Hadlock's formula[3] – which includes measurements of biparietal diameter (BPD), abdominal circumference (AC) and femur length (FL) – has been shown to provide the best accuracy[4] and is thus frequently used. Two-dimensional ultrasound (2D-US) is routinely used to assess fetal biometric parameters; however, the variability caused by observers' random errors is considered to be the main limitation of this method[5,6]. The observer's experience, maternal biotype and fetal position are factors that significantly influence the acquisition of ideal fetal planes/images in order to perform such measurements reliably.

Three-dimensional ultrasound (3D-US) has been proven to be useful and reliable for measurements in gynecology and obstetrics (e.g. endometrial thickness[7], lower uterine segment[8] and even fetal biometry[9,10]). The 3D-US multiplanar mode allows the simultaneous view of three orthogonal planes and the volume acquired can be rotated and centralized to provide a standardized multiplanar view (SMV)[7]. This, at least theoretically, allows the

---

Correspondence to: Dr W. P. Martins, Departamento de Ginecologia e Obstetrícia da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes, 3900, 8 andar., Ribeirão Preto, São Paulo, Brazil, CEP: 14049-900
(e-mail: wpmartins@gmail.com)

Accepted: 2 November 2011

operator to perform more reliable measurements because it reduces the risk of performing oblique or off-center measurements. With technological advances and the training of professionals, the use of 3D-US has spread widely, which makes the method more feasible for implementation in prenatal care; particularly in cases when even small errors might affect clinical management.

The purpose of this study was to evaluate and compare the intra- and interobserver reliability and agreement for fetal measurements (BPD, AC and FL) and EFW in singleton pregnant women between 24 and 40 weeks, assessed by 2D-US and 3D-US in different and blinded evaluations.

METHODS

The following terminologies were used in this study. *Systematic error*: a predictable error usually leading to underestimated or overestimated measures[11]. It may occur when there is something wrong with the method, or because the method is not being used correctly by the observer. *Random error*: caused by unpredictable changes when performing the measurement and can be revealed by repeating the measurements[11]; it is closely related to the concept of precision. Random errors often have a normal distribution scattered about the true value. *Reliability*: relates the magnitude of the measurement error in observed measurements to the inherent variability between subjects[12]. *Agreement*: quantifies how close two measurements made on the same subject are independently of the variability between subjects[12]. *Validity*: refers to differences observed between the recorded value from one observer and the 'true' value that should be estimated by a suitable reference standard[13] (e.g. EFW by ultrasound and weight at birth). Validity was not assessed in this study.

The sample size was calculated for a 33% reduction in the magnitude of the SD of the difference among the measures (e.g. a reduction of 75g[14] to 50g). If this reduction occurs, the interval in which 68% of differences were found (± 1.0 SD) will contain 87% of the differences (± 1.5 SD). Therefore, we would need to evaluate approximately 100 pregnant women to have a 90% power to demonstrate this difference. The choice of 33% reduction was arbitrary and based on the authors' perception of clinical relevance.

We invited pregnant women undergoing obstetric ultrasound at Escola de Ultra-sonografia e Reciclagem Médica de Ribeirão Preto (EURP) to participate in the study. Inclusion criteria were: known date of last menstrual period (LMP); gestational age between 24 and 40 weeks according to LMP; ultrasound examination performed before 14 weeks, with gestational age determined by crown–rump length (CRL) to be within 5 days of the gestational age based on the LMP (the latter was considered as the 'true' gestational age); singleton pregnancies; and agreement to participate in this study, providing signed informed consent. Exclusion criteria were: fetal malformation that prevented biometric measurement and failure to undergo three consecutive ultrasound examinations. This study was approved by both local Institutional Review Boards.

Two observers, both specialists in obstetrics and gynecology, conducted the ultrasound examinations. The ultrasound scans followed this sequence: Observer 1 (J.C.L.: 6 years of experience with 2D-US and 3 years of experience with 3D-US) measured the fetuses using conventional 2D-US and acquired 3D datasets for later evaluation. When Observer 1 had completed the evaluation he left the examination room and Observer 2 (A.H.M.: with 3 years of experience with 2D-US, and 1 year of experience with 3D-US) entered to perform the same steps for the determination of interobserver reliability and agreement. Afterwards, Observer 1 re-entered the room and repeated the same steps for the evaluation of intraobserver reliability and agreement.

The two observers were blinded to the measurements by a label placed on the numeric display. Each evaluation took approximately 20 minutes, resulting in a total evaluation time of approximately 1 hour for each subject. We used a Voluson 730 Expert (GE Healthcare Ultrasound, Milwaukee, WI, USA) ultrasound machine with a 4–8-MHz transabdominal probe (RAB 4–8L). The flow chart of the study design is presented in Figure 1. In each of the three assessments, the observer performed measurements for the following biometric parameters: BPD, FL, anteroposterior abdominal diameter (APAD) and transverse abdominal diameter (TAD) (Figure 2). The determination of AC was performed using the following formula: $AC = (APAD + TAD) \times 1.57$[15].

Measurements were performed as follows[16]. BPD was measured at the level of the thalami and cavum septi pellucidi. The cerebellar hemispheres should not be visible in this scanning plane. The measurement was taken from the outer edge of the proximal skull to the inner edge of the distal skull. APAD and TAD were measured from the skin line on a transverse view at the level of the junction of the umbilical vein, portal sinus and fetal stomach, when visible. For FL, the long axis of the femoral shaft was measured with the beam of insonation being perpendicular to the shaft, excluding the distal femoral epiphysis.

During each evaluation the observer acquired three, 3D-US datasets: head, abdomen and thigh. The acquisition of the 3D-US datasets was performed by insonating the fetus in the same way as when measuring BPD (for head volume), AC (for abdomen volume) and FL (thigh volume) in 2D-US. The same biometric measurements were assessed in the 3D-US datasets by the observers in a blinded manner using the software 4D View 10.0 (GE Healthcare Ultrasound) on their personal computers. Fetal measurements were performed in exactly the same way as for 2D-US, but using the SMV[7]. The SMVs of the head, abdomen and thigh were obtained, aiming to align the fetal parts properly before making the measurements in the A-plane (Figure 3).

EFW was determinted, on 2D-US and 3D-US, using the formula[3]: $\text{Log}_{10}\ EFW = 1.335 - 0.0034 \times AC \times FL + 0.0316 \times BPD + 0.0457 \times AC + 0.1623 \times FL$.

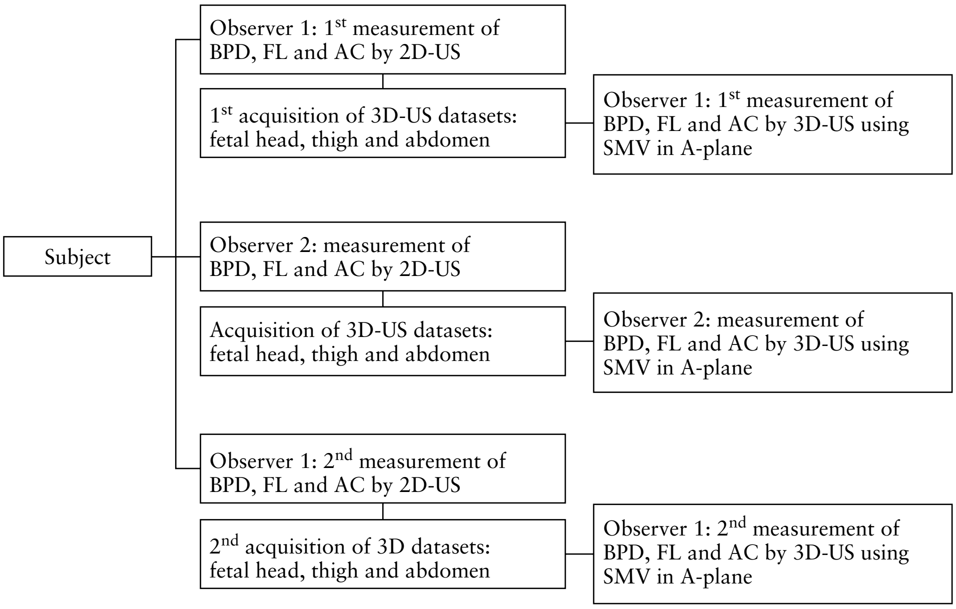

Figure 1 Flow chart of the study design. Abdominal circumference (AC) was determined by measurements of the two abdominal diameters and estimated fetal weight (EFW) was assessed using Hadlock's formula[3]. 2D-US, two-dimensional ultrasound; 3D-US, three-dimensional ultrasound; BPD, biparietal diameter; FL, femur length; SMV, standardized multiplanar view.

Figure 2 Measurements performed by two-dimensional ultrasound (2D-US): (a) fetal head and measurement of biparietal diameter; (b) fetal abdomen and measurement of anteroposterior and transverse abdominal diameters; and (c) fetal thigh and measurement of femur length.

Statistical analysis

Data were analyzed using the programs PASW 18.0 for Windows (SPSS Inc., Chicago, IL, USA) and GraphPad 5.0 for Windows (GraphPad Software, San Diego, CA, USA). After assessing normality of distribution using the Kolmogorov–Smirnov test, we determined the mean, SD and minimum and maximum values of the parameters studied.

We evaluated the intra- and interobserver reliability using the intraclass correlation coefficient (ICC, two-way mixed, single measures) and its 95% CI.

We examined the occurrence of systematic errors (measurements on average higher or lower) between 2D-US and 3D-US by paired *t*-tests using data from the first set of measurements of Observer 1. We did not repeat the same comparisons using other data because we wanted to avoid the increased risk of type 1 error that occurs when performing multiple tests for the same purpose.

We calculated the SD of the signed absolute and relative intra- and interobserver differences between measurements for both 2D-US and 3D-US. The SD calculated from the differences observed between two sets of measurements using the same method indicates the variation related to random measurement errors because the measurements were performed in the same fetuses and in a short period of time. The following differences were calculated.

Intraobserver absolute difference = (2nd measurement by Observer 1) − (1st measurement by Observer 1)

Interobserver absolute difference = (Measurement by Observer 2) − (1st measurement by Observer 1)

Intraobserver relative difference = ((2nd measurement by Observer 1) − (1st measurement by Observer 1))/(Average value between these measurements)

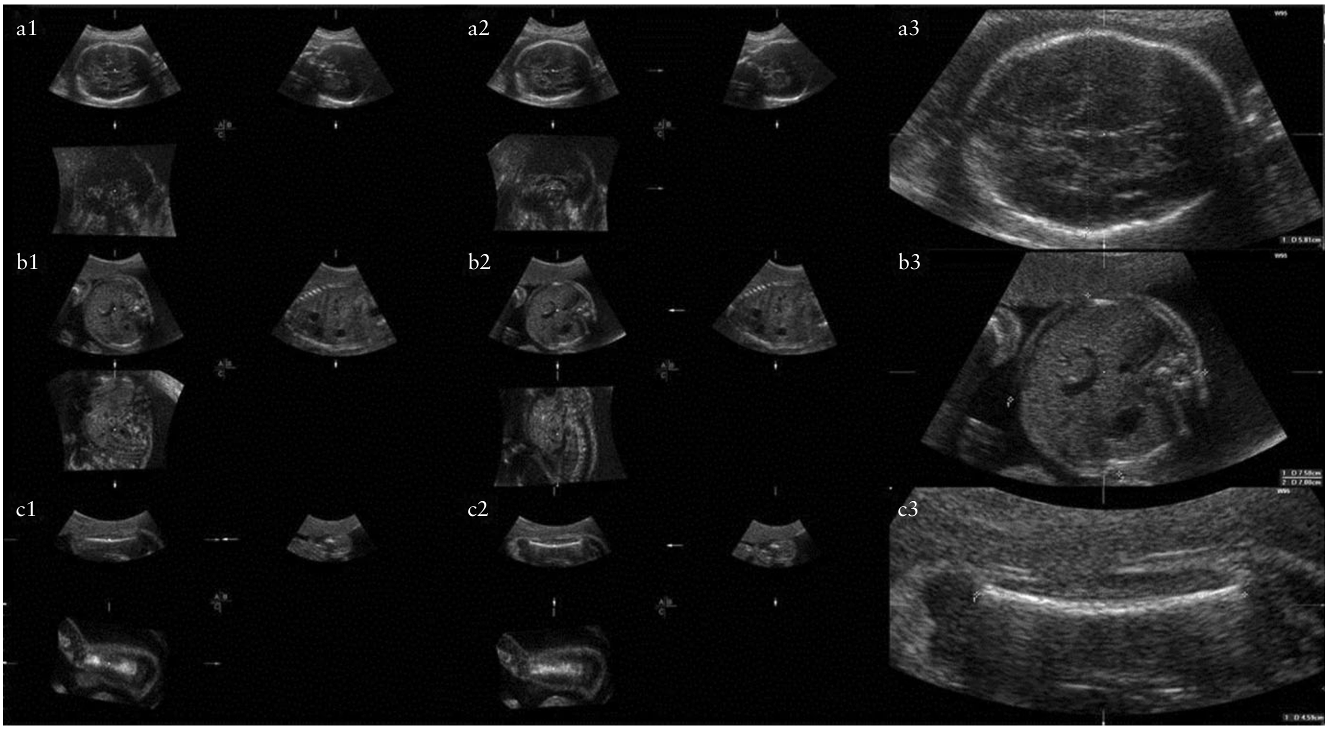

Figure 3 Measurements performed by three-dimensional ultrasound (3D-US): (a1) fetal head as acquired in unmodified multiplanar view (UMV); (a2) fetal head, after adjustments, in standardized multiplanar view (SMV); (a3) measurement of biparietal diameter using magnified A-plane of SMV in full screen; (b1) fetal abdomen in UMV; (b2) fetal abdomen in SMV; (b3) measurement of abdominal diameters using magnified A-plane of SMV in full screen; (c1) fetal thigh in UMV; (c2) fetal thigh in SMV; and (c3) measurement of femur length using magnified A-plane of SMV in full screen.

Interobserver relative difference = ((Measurement by Observer 2) − (1^{st} measurement by Observer 1))/(Average value between these measurements)

The SD values calculated from these differences were then compared between 2D-US and 3D-US using Pitman's test for correlated variances[17,18]. In order to make the meaning of SD more intuitive, we presented data as the interval expected to comprise 95% of random measurement errors (± 1.96 SD). Intra- and interobserver errors using 2D-US and 3D-US were also examined visually using Bland–Altman plots.

Additionally, we examined agreement by comparing the proportions of absolute differences that were below defined cut-off points. For this purpose we used McNemar's test, a non-parametric method used on nominal data that can be applied to 2 × 2 contingency tables with a dichotomous trait and matched pairs of subjects[19]. We used two sets of cut-off points: one set was based on the authors' perception of trivial error (BPD and FL ≤ 0.15 cm, AC ≤ 1.50 cm and EFW ≤ 200 g and $\leq 10\%$); and the other set was equal to 1.0 SD obtained from the intraobserver measurement differences of 2D-US, which should account for approximately 68% of these differences. We considered $P < 0.05$ as statistically significant.

RESULTS

We included 105 pregnant women, but three were subsequently excluded because they declined to complete the sequence of three examinations. Therefore, 102 singleton pregnancies were completely evaluated between 25 + 0 and 38 + 3 weeks' gestational age (mean ± SD = 30.2 ± 3.3 weeks): 10 pregnancies from 24 + 0 to 25 + 6 weeks, 16 pregnancies from 26 + 0 to 27 + 6 weeks, 28 pregnancies from 28 + 0 to 29 + 6 weeks, 15 pregnancies from 30 + 0 to 31 + 6 weeks, 16 pregnancies from 32 + 0 to 33 + 6 weeks, 12 pregnancies from 34 + 0 to 35 + 6 weeks, three pregnancies from 36 + 0 to 37 + 6 weeks and two pregnancies from 38 + 0 to 40 + 0 weeks. Complete evaluation of all 102 pregnant women resulted in 306 measurements of BPD, AC, FL and EFW by 2D-US (three per subject). In total, 918 3D-US datasets were acquired (nine per subject; three sets of three, 3D-US datasets: head, abdomen and thigh) and the same 306 measurements of BPD, AC, FL and EFW were performed using 3D-US. Table 1 shows parameters for the distribution of BPD, FL, AC and EFW.

We did not observe any significant systematic error between 2D-US and 3D-US: BPD, 7.42 ± 0.78 vs. 7.41 ± 0.80, $P = 0.65$; FL, 5.61 ± 0.71 vs. 5.58 ± 0.71, $P = 0.21$; AC, 25.45 ± 3.63 vs. 25.43 ± 3.60, $P = 0.82$; and EFW, 1516 ± 589 vs. 1509 ± 581, $P = 0.40$ (2D-US *vs* 3D-US respectively, P-values were obtained from paired t-tests using data from the first set of measurements of Observer 1).

Table 1 Fetal measurements obtained by two observers using two-dimensional (2D-US) and three-dimensional (3D-US) ultrasound

Variable	*2D-US*			*3D-US*		
	Observer 1 (1st measurement)	*Observer 1 (2nd measurement)*	*Observer 2*	*Observer 1 (1st measurement)*	*Observer 1 (2nd measurement)*	*Observer 2*
BPD (cm)	7.42 ± 0.78 (5.55–9.50)	7.45 ± 0.81 (5.68–9.76)	7.44 ± 0.79 (5.73–9.43)	7.41 ± 0.80 (5.35–9.49)	7.41 ± 0.79 (5.41–9.44)	7.42 ± 0.79 (5.53–9.46)
FL (cm)	5.61 ± 0.71 (4.26–7.33)	5.60 ± 0.71 (4.27–7.32)	5.59 ± 0.71 (4.25–7.34)	5.58 ± 0.71 (4.29–7.24)	5.59 ± 0.72 (4.29–7.23)	5.58 ± 0.71 (4.32–7.47)
AC (cm)	25.45 ± 3.63 (18.18–34.45)	25.36 ± 3.56 (18.24–35.56)	25.64 ± 3.50 (18.90–33.63)	25.43 ± 3.60 (17.93–34.41)	25.37 ± 3.48 (18.64–35.88)	25.37 ± 3.53 (18.42–34.48)
EFW (g)	1516 ± 589 (599–3332)	1511 ± 583 (599–3161)	1530 ± 583 (626–3209)	1509 ± 581 (600–3371)	1502 ± 571 (635–3621)	1502 ± 577 (636–3372)

Data are shown as mean ± SD (range). AC, abdominal circumference; BPD, biparietal diameter; EFW, estimated fetal weight; FL, femur length.

When assessing reliability (Table 2), we observed that the ICC values for 3D-US were higher than the values observed for 2D-US for both intra- and interobserver evaluations: however, only for FL was there no overlap in the 95% CI.

When visually examining the Bland–Altman plots and LoA (Figures 4 and 5), we observed that intra- and interobserver differences were evenly distributed above and below zero, suggesting no systematic error within or between observers. We also observed that the LoAs were wider for 2D-US, suggesting that random errors were smaller when using 3D-US. When comparing intra- and interobserver random measurement errors, for both absolute and relative differences, we observed that the SD of differences was smaller for 3D-US compared with 2D-US, and that therefore the expected agreement would be better (Table 3). Statistical significance, examined using Pitman's test, was observed in all comparisons, except for the intraobserver relative differences when measuring AC ($P = 0.07$).

We also observed that 3D-US had a higher proportion of differences below the arbitrarily defined cut-off points (Table 4). Statistical significance, examined using McNemar's test, was observed in all comparisons, except when comparing the proportions of intraobserver differences for EFW $\leq 10\%$ ($P = 0.12$).

Table 2 Evaluation of intraobserver and interobserver reliabilities of fetal measurements made using two-dimensional (2D-US) and three-dimensional (3D-US) ultrasound

Fetal measurement	*2D-US*	*3D-US*
BPD		
Intraobserver	0.982 (0.973–0.988)	0.989 (0.983–0.992)
Interobserver	0.975 (0.963–0.983)	0.988 (0.982–0.992)
FL		
Intraobserver	0.989 (0.984–0.993)	0.996 (0.995–0.998)
Interobserver	0.987 (0.981–0.991)	0.996 (0.993–0.997)
AC		
Intraobserver	0.944 (0.918–0.962)	0.964 (0.947–0.976)
Interobserver	0.933 (0.903–0.954)	0.962 (0.944–0.974)
EFW		
Intraobserver	0.974 (0.962–0.982)	0.987 (0.981–0.991)
Interobserver	0.973 (0.961–0.982)	0.985 (0.978–0.990)

Data are given as intraclass correlation coefficient (95% CI). AC, abdominal circumference; BPD, biparietal diameter; EFW, estimated fetal weight; FL, femur length.

DISCUSSION

We observed that 3D-US was associated with improvement in the reliability and agreement in all evaluated parameters: ICCs were higher, especially for FL; SD values were smaller; higher proportions of differences were below the arbitrarily defined cut-off points; and narrower LoA were observed. All these findings suggest that measurements performed with 3D-US are less influenced by random errors[20].

Two previous studies have compared 2D-US and 3D-US for fetal biometry measurements. One of these studies evaluated the intra- and interobserver agreement (three pairs of doctors), when examining 36 fetuses (12 per pair of doctors) at 24–32 weeks[9]. Each pair of doctors measured BPD, head circumference (HC), AC and FL: the authors observed that 3D-US has the potential to reduce intra- and interobserver measurement variation, especially in FL measurements. The other study evaluated the intraobserver reliability and agreement for BPD, HC, AC and FL measurements in 50 fetuses performed by an inexperienced observer using 2D-US and 3D-US, which were compared with the results of 2D-US measurements performed by an experienced observer[10]: both 2D-US and 3D-US measurements were reproducible and showed good agreement with those obtained by an experienced operator; additionally, the use of 3D-US by an inexperienced operator allows faster measurements than 2D-US ultrasound and also seems to facilitate the acquisition of higher-quality images for measurement of AC. However, these two studies have an important limitation: both evaluated the reproducibility of 3D-US using a single acquired dataset (only one acquisition), while 2D-US measurements were performed at different times, allowing fetal movements. Additionally, the studies did not analyze the effect of the improvement in the reliability and agreement of the measures upon

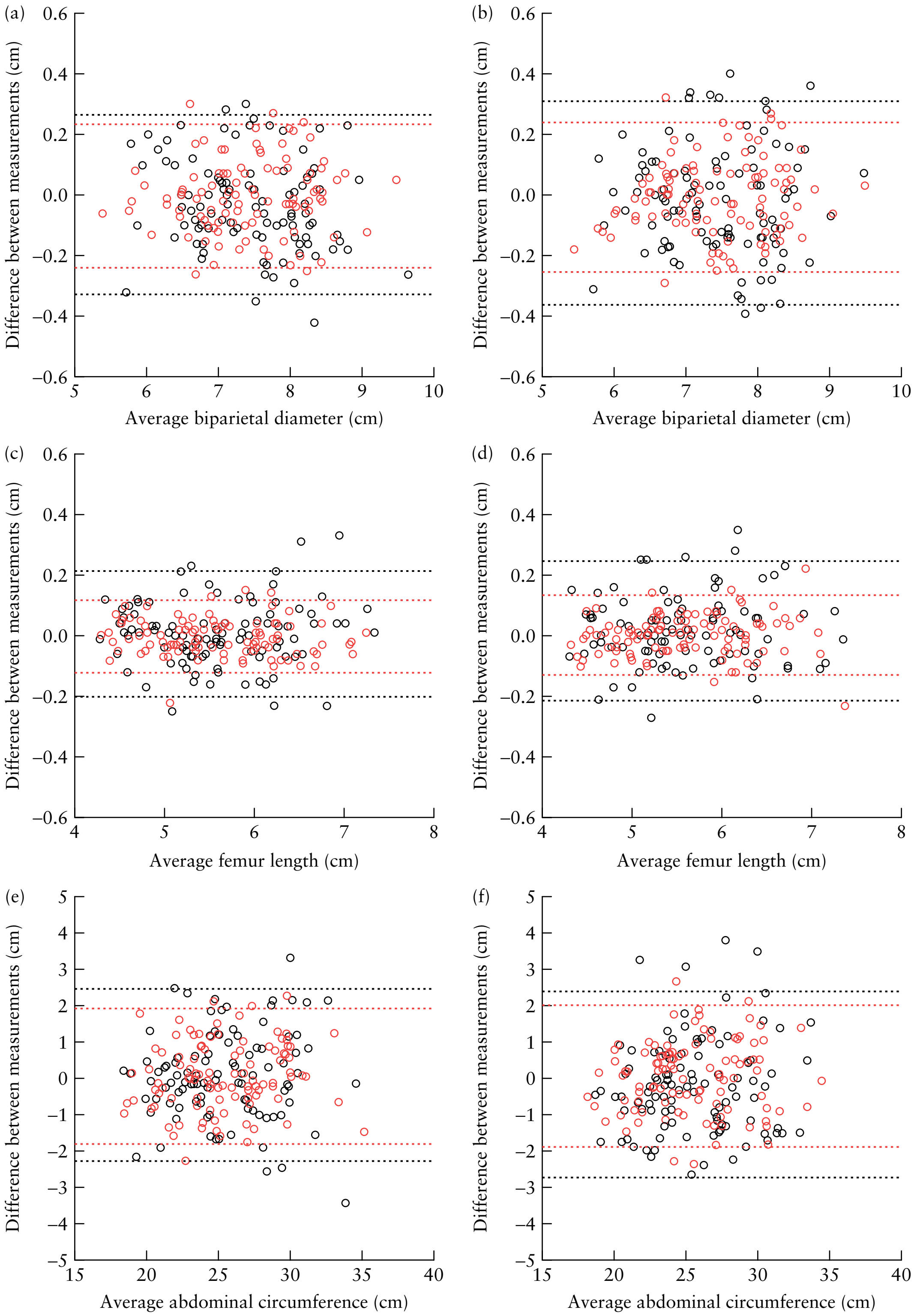

Figure 4 Bland–Altman plots showing limits of agreements (dotted lines) fot intraobserver (a, c, e) and interobserver (b, d, f) differences when measuring biparietal diameter (a, b), femur length (c, d) and abdominal circumference (e, f) by two-dimensional (black) and three-dimensional (red) ultrasound.

EFW, and neither study included pregnancy in the late third trimester: the time periods were between 24 and 32 weeks[9], and between 17 and 34 weeks[10]. In clinical practice, the reproducibility of these measurements is even more relevant near term, when pathological changes in intrauterine growth are more prevalent as a result of placental and nutritional dysfunctions[21,22].

Some aspects of this study enabled a more accurate evaluation of the reliability and agreement of fetal biometric measurements: 102 pregnant women were

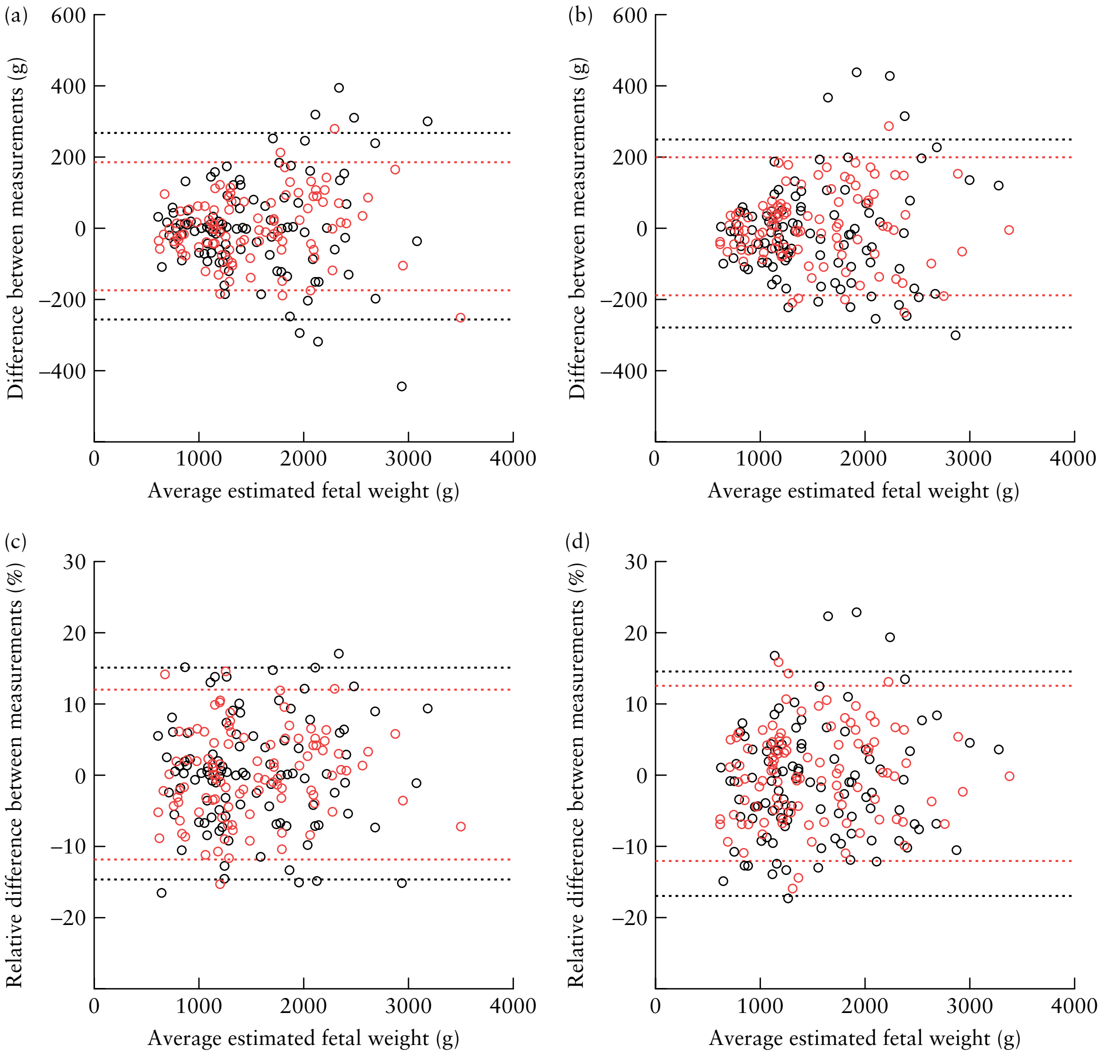

Figure 5 Bland–Altman plots showing limits of agreements (dotted lines) for intraobserver (a, c) and interobserver (b, d) absolute (a, b) and relative (c, d) differences in estimated fetal weight using measurements performed by two-dimensional (black) and three-dimensional (red) ultrasound.

Table 3 Comparison of intervals expected to comprise 95% of random intraobserver and interobserver errors between two-dimensional (2D-US) and three-dimensional (3D-US) ultrasound

	Intraobserver			*Interobserver*		
	2D-US	*3D-US*	P	*2D-US*	*3D-US*	P
Signed absolute differences						
BPD (cm)	± 0.29	± 0.24	0.03	± 0.34	± 0.25	< 0.01
FL (cm)	± 0.21	± 0.12	< 0.01	± 0.22	± 0.13	< 0.01
AC (cm)	± 2.36	± 1.86	0.02	± 2.55	± 1.94	0.01
EFW (g)	± 262.1	± 181.0	< 0.01	± 265.0	± 194.5	< 0.01
Signed relative differences						
BPD (%)	± 4.01	± 3.17	0.02	± 4.60	± 3.33	< 0.01
FL (%)	± 3.63	± 2.16	< 0.01	± 3.99	± 2.27	< 0.01
AC (%)	± 8.99	± 7.52	0.07	± 9.98	± 7.59	0.01
EFW (%)	± 14.82	± 11.96	0.03	± 15.72	± 12.33	0.01

Signed absolute difference = (Measurement B) – (Measurement A). Signed relative difference = ((Measurement B) – (Measurement A))/(Average value of these measurements). Interval expected to comprise 95% of random measurement errors = ± 1.96 SD determined from the differences between measurements; *P*-values determined using Pitman's test for correlated variances. AC, abdominal circumference; BPD, biparietal diameter; EFW, estimated fetal weight; FL, femur length.

evaluated three times, giving a total of 306 ultrasound examinations and 918 3D-US datasets, which allowed enough power to detect improvement in random error > 30%. Furthermore, the study included assessment of the reproducibility of EFW, blinded measurements and the analysis of different 3D-US datasets for the assessment of intra- and interobserver reliability and agreement. The latter point is important because when considering 3D-US reproducibility we have to consider both the variation that may occur because of errors in the implementation of the measures (reproducibility of analysis) and the variations that may occur as a result of fetal motion, which will only be assessed if the acquisitions are made at different times (reproducibility of acquisition)[23,24]. This is particularly important for reproducibility of the measurement of AC because it is more likely to be affected by fetal movements. On the other hand, the small number of observers (only two) may be considered as a limitation of the present study.

The better reliability and agreement observed when using 3D-US for fetal measurements and EFW should

Table 4 Proportions of differences below arbitrary cut-off values for intraobserver and interobserver differences of fetal measurements made using two-dimensional (2D-US) and three-dimensional (3D-US) ultrasound

	Intraobserver			*Interobserver*		
Cut-off points	*2D-US*	*3D-US*	P	*2D-US*	*3D-US*	P
Author*						
BPD $\leq$ 0.15 cm	66.67	79.41	0.04	62.75	77.45	0.02
FL $\leq$ 0.5 cm	86.27	99.02	< 0.01	83.33	98.04	< 0.01
AC $\leq$ 1.50 cm	76.47	89.22	0.02	74.51	89.22	0.01
EFW $\leq$ 200 g	89.22	97.06	0.02	86.27	97.06	< 0.01
EFW $\leq$ 10%	80.39	89.22	0.12	78.43	91.18	0.02
1.0 SD†						
BPD $\leq$ 0.15 cm	66.67	79.41	0.04	62.75	77.45	0.02
FL $\leq$ 0.11 cm	74.51	95.10	< 0.01	74.51	93.14	< 0.01
AC $\leq$ 1.20 cm	71.57	83.33	0.04	62.75	78.43	0.01
EFW $\leq$ 134 g	72.55	86.27	0.02	68.63	80.39	0.04
EFW $\leq$ 7.6%	69.61	85.29	0.01	65.69	81.37	0.01

Data expressed as percent. *P*-values obtained using McNemar's test; non-signed differences (modulus) used to calculate proportions. *Cut-off determined by authors' perception of trivial error. †Cut-off equal to 1.0 SD of intraobserver differences obtained by 2D-US. AC, abdominal circumference; BPD, biparietal diameter; EFW, estimated fetal weight; FL, femur length.

be considered by clinicians. Although the improvement was not huge, the observed difference might be important in some obstetric situations, when there is a need for serial evaluation of fetal growth. However, more studies are still needed, preferably including situations where precise measurements are more important, such as in cases of intrauterine growth restriction, macrosomic fetuses, evaluation of weekly weight gain and twins, particularly when there is a suspicion of weight discordance. Other studies should also consider including a greater number of observers, preferably with different levels of training. Only after all of these studies have been performed will we have a clearer idea of the magnitude of the improvement provided by 3D-US in the most relevant situations.

REFERENCES

1. Nahum GG, Stanislaw H. Ultrasonographic prediction of term birth weight: how accurate is it? *Am J Obstet Gynecol* 2003; **188**: 566–574.
2. Pates JA, McIntire DD, Casey BM, Leveno KJ. Predicting macrosomia. *J Ultrasound Med* 2008; **27**: 39–43.
3. Hadlock FP, Harrist RB, Sharman RS, Deter RL, Park SK. Estimation of fetal weight with the use of head, body, and femur measurements – a prospective study. *Am J Obstet Gynecol* 1985; **151**: 333–337.
4. Burd I, Srinivas S, Pare E, Dharan V, Wang E. Is sonographic assessment of fetal weight influenced by formula selection? *J Ultrasound Med* 2009; **28**: 1019–1024.
5. Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. *Ultrasound Obstet Gynecol* 2005; **25**: 80–89.
6. 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.
7. Martins WP, Raine-Fenning NJ, Leite SP, Ferriani RA, Nastri CO. A standardized measurement technique may improve the reliability of measurements of endometrial thickness and volume. *Ultrasound Obstet Gynecol* 2011; **38**: 107–115.
8. Martins WP, Barra DA, Gallarreta FM, Nastri CO, Filho FM. Lower uterine segment thickness measurement in pregnant women with previous Cesarean section: reliability analysis using two- and three-dimensional transabdominal and transvaginal ultrasound. *Ultrasound Obstet Gynecol* 2009; **33**: 301–306.
9. Chan LW, Fung TY, Leung TY, Sahota DS, Lau TK. Volumetric (3D) imaging reduces inter- and intraobserver variation of fetal biometry measurements. *Ultrasound Obstet Gynecol* 2009; **33**: 447–452.
10. Yang F, Leung KY, Lee YP, Chan HY, Tang MH. Fetal biometry by an inexperienced operator using two- and three-dimensional ultrasound. *Ultrasound Obstet Gynecol* 2010; **35**: 566–571.
11. Taylor JR. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* (2nd edn). University Science Books: Sausalito, 1997.
12. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol* 2008; **31**: 466–475.
13. Khan KS, Chien PF. Evaluation of a clinical test. I: assessment of reliability. *BJOG* 2001; **108**: 562–567.
14. 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.
15. Campbell S, Wilkin D. Ultrasonic measurement of fetal abdomen circumference in the estimation of fetal weight. *Br J Obstet Gynaecol* 1975; **82**: 689–697.
16. Salomon LJ, Alfirevic Z, Berghella V, Bilardo C, Hernandez-Andrade E, Johnsen SL, Kalache K, Leung KY, Malinger G, Munoz H, Prefumo F, Toi A, Lee W. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obstet Gyneol* 2011; **37**: 116–216.
17. Pitman EJG. A note on normal correlation. *Biometrika* 1939; **31**: 9–12.
18. Bartko JJ. Measures of agreement: a single procedure. *Stat Med* 1994; **13**: 737–745.
19. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12**: 153–157.
20. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, Roberts C, Shoukri M, Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; **64**: 96–106.
21. Cox P, Marton T. Pathological assessment of intrauterine growth restriction. *Best Pract Res Clin Obstet Gynaecol* 2009; **23**: 751–764.
22. Marsal K. Obstetric management of intrauterine growth restriction. *Best Pract Res Clin Obstet Gynaecol* 2009; **23**: 857–870.
23. Martins WP. Three-dimensional power Doppler: validity and reliability. *Ultrasound Obstet Gynecol* 2010; **36**: 530–533.
24. Martins WP, Raine-Fenning NJ. Analysis and acquisition reproducibility of 3D power Doppler. *Ultrasound Obstet Gynecol* 2010; **36**: 392–393; author reply 393–394.