Diagnostic and Interventional Imaging 101 (2020) 481–487

Original article/*Obstetric imaging*

# Prospective assessment of reproducibility of three-dimensional ultrasound for fetal biometry

G. Ambroise Grandjean[a,b,c,*], P. Berveiller[d,e], G. Hossu[f], P. Noble[g], M. Chamagne[b], O. Morel[a,b]

[a] *Inserm, IADI, Université de Lorraine, 54000 Nancy, France*
[b] *Department of Obstetrics and Gynecology, Centre hospitalier regional universitaire de Nancy, 54000 Nancy, France*
[c] *Midwifery Department, Université de Lorraine, 54000 Nancy, France*
[d] *Department of Obstetrics and Gynecology, Centre hospitalier intercommunal de Poissy Saint-Germain-en-Laye, 78300 Poissy, France*
[e] *Université Versailles-Saint-Quentin, 78180 Montigny-le-Bretonneux, France*
[f] *CIC-IT, Centre hospitalier regional universitaire de Nancy, 54000 Nancy, France*
[g] *Department of Obstetrics and Gynecology, Port-Royal, hôpital Cochin, Assistance Publique–Hôpitaux de Paris, 75014 Paris, France*

ARTICLE INFO

*Keywords:*
Gestational age
Reproducibility of results
Ultrasonography,Prenatal
Three-dimensional ultrasound

ABSTRACT

*Purpose:* To compare fetal ultrasound measurements performed by two observers with different levels of experience and evaluate the potential contribution of the use of three-dimensional (3D) ultrasound on repeatability, reproducibility and agreement of two-dimensional (2D) and 3D-derived measurements.
*Materials and methods:* Two observers (one senior and one junior) measured head circumference (HC), abdominal circumference (AC) and femur length (FL) in 33 fetuses (20 to 40 weeks of gestation). Each observer performed two series of 2D measurements and two series of 3D measurements (*i.e.*, measurements derived from triplane volume processing). Measurements were converted into Z-scores according to gestational age. Variability between the different series of measurements was studied using Bland–Altmann plots and intra-class correlation coefficients (ICC).
*Results:* Agreement with the 2D measurements of the senior observer was higher in 3D than in 2D for the junior observer (systematic differences of −0.4, −0.2 and −0.8 Z-score vs. −0.1, −0.1 and −0.6 for HC, AC and FL on 2D and 3D datasets, respectively). The use of 3D ultrasound improved junior observer repeatability (ICC = 0.94, 0.88, 0.90 *vs.* 0.94, 0.94 and 0.96 for HC, AC and FL in 2D and 3D, respectively). The reproducibility was greater using the junior observer 3D datasets (ICC = 0.75, 0.60 and 0.45 *vs.* 0.79, 0.89 and 0.63 for HC, AC and FL, respectively).
*Conclusion:* The use of 3D ultrasound improves the consistency of the measurements performed by a junior observer and increases the overall repeatability and reproducibility of measurements performed by observers with different levels of experience.

## 1. Introduction

Screening for fetal growth abnormalities relies primarily on accurate biometric measurements performed on two-dimensional (2D) ultrasound images, allowing the estimation of fetal weight [1]. As with all screening tests, it is essential to maintain strict control over the reproducibility of the measurements used [2,3]. The operator-dependent nature of ultrasound compared with other imaging techniques is well understood [4] and impacts negatively on the quality of the measurements. The variability of ultrasound measurements is difficult to quantify probably due to several factors including the identification of anatomical landmarks, plane acquisition and caliper positioning [3]. However, even if this variability is marginal and acceptable for experienced operators, it remains substantial for inadequately trained operators [4–7].

The use of three-dimensional (3D) ultrasound allows a potential pathway to improve the accuracy of fetal measurements. Offline analysis of stored 3D volumes to select the plane, which corresponds to the conventional 2D ultrasound criteria, can be performed by identifying anatomical landmarks without specific skill

*Abbreviations:* 2D, Two-Dimensional; 3D, Three-Dimensional; AC, Abdominal Circumference; BMI, Body Mass Index; CI, Confidence Interval; EFW, Estimated Fetal Weight; FL, Femoral Length; HC, Head Circumference; HSD, Honestly Significant Difference; ICC, Intraclass Correlation Coefficients; WG, Weeks' Gestation.

* Corresponding author at: Inserm, IADI, Université de Lorraine, 54000 Nancy, France.

*E-mail address:* g.ambroise@chru-nancy.fr (G. Ambroise Grandjean).

https://doi.org/10.1016/j.diii.2020.03.004
2211-5684/

at manipulating the probe [8]. It was previously demonstrated that repeatability and operator reproducibility could be improved by using 3D ultrasound volumes to perform biometric measurements [9,10]. Improvement might be especially significant for inexperienced operators [11]. The use of 3D ultrasound was also shown to reduce the time needed to perform the ultrasound and to lead to an improved image quality.

However, in previously published reports, fetal biometric measurements were performed at different gestational ages, which did not allow for the direct comparison of data. Indeed, expressed in millimeters, the observed variability between repeated measurements of the same parameter increases with increasing gestational age. By using the Z-score, this difficulty can be overcome [12].

Although 3D ultrasound has an a priori lower risk of variability than 2D ultrasound, it still presents a risk of intra- and inter-operator variability for two reasons: firstly, due to the quality of the acquisition of the volume [8] and secondly due to the quality of the subsequent analysis (selection of the correct plane and caliper placement). The specific impact of this second step (*i.e.*, processing volumes) has not yet been assessed.

The primary goal of this pilot study was to compare fetal ultrasound measurements performed by two observers of different levels of experience and evaluate the potential contribution of 3D ultrasound on repeatability, reproducibility and agreement of 2D and 3D-derived measurements. The additional study had the aim of assessing the impact of the observer's experience on the processing of the ultrasound volumes.

## 2. Materials and methods

### 2.1. Patients

Data were collected during a prospective cross-sectional pilot study undertaken on women referred for ultrasound examination to an academic hospital. The standard management of pregnant women in our institution includes an ultrasound examination performed by an expert sonographer (with findings being reported in the patient's medical notes). An optional additional ultrasound examination may be performed by a student as part of training. Oral consent is obtained from the patient before the additional scan is undertaken. This additional scan results in a slightly longer ultrasound examination. Data acquired during this scan do not contain patient information or result in changes to the patient's care.

This study was performed from November 1st, 2015 until the January 31st, 2016. Women referred for an ultrasound examination involving biometric measurements between 17 and 41 weeks' gestation (WG) were invited to participate. The non-inclusion criterion was the refusal to participate in the optional part of the examination. The exclusion criterion was the presence or suspicion of a fetal abnormality.

### 2.2. Ultrasound protocol and image analysis

The measurements were performed on a Voluson E8® (GE Healthcare) using a 4–8 MHz probe. During the ultrasound examinations, a mask was placed on the screen to prevent the observer from seeing the measurements and gestational age-related percentiles. Each 2D and 3D measurement was performed twice by each observer to evaluate the reproducibility.

All 2D and 3D biometric measurements were performed according to the methodology published in the Intergrowth 21 study [1]. The procedures for analyzing the volumes were developed and implemented by adapting procedures from previous publications and using the co-author's expertise [8]. The 3D volume acquisition was obtained in triplane mode with the sweep angle set at 60°. This angle had been previously defined to encompass all the relevant structures in a fetus at 41 WG. In order to limit the distortion of the volumes, the acquisition planes were: a transverse transthalamic plane, a transverse abdominal plane and a plane allowing the visualization of the uppermost femur (with less than 40° angulation compared to the horizontal) for the volumes of the cephalic pole, the abdomen and the thigh respectively [8]. For very active fetuses, the speed of acquisition was increased. The pressure of the probe on the maternal abdomen was adjusted in order not to deform the contours of the fetal anatomy. All of these acquisition techniques, as well as the visual assessment of the absence of artifact caused by fetal movement during the volume acquisition, were covered in the teaching received by the junior observer before the study.

The time taken to obtain the 2D planes and 3D volumes, as well as the analysis of the volumes, were recorded. For 2D measurement procedures and 3D volumes acquisition, the observers kept an empty screenshot at the beginning and end of each process to calculate the duration. For volume processing procedures, the observers used an external timer.

### 2.3. Data collection and definitions

Data were collected during examinations performed successively by both a senior observer and a junior observer. The senior observer (G. A.) was an expert sonographer and a trainer for fetal biometry training programs. This observer has conducted over 2000 ultrasound examinations, with both 2D and 3D technique, over the last five years. The junior observer (M.C.) was a trainee in obstetrics and gynecology and had performed less than 50 2D ultrasound examinations in the course of his training and had minimal experience in 3D ultrasound. Before the study commenced, the junior observer received two hours of theoretical training and four hours of hands-on training on how to perform 2D and 3D fetal biometry including measurement methodology, anatomical landmarks, volumes acquisition and 3D volume processing.

Conventional 2D measurements performed by the senior observer were chosen as the reference standard. The index tests assessed were 2D measurements performed by the junior observer and 3D measurements performed both by the senior and the junior observers. The STARD standards were used to report the results [13].

For each fetus, datasets comprising of the head circumference (HC), the abdominal circumference (AC) and the femur length (FL) were recorded twice by the two observers and further referred to as datasets 2Dsenior and 2Djunior. Similarly, 3D volumes (cephalic pole, abdomen, and thigh) were acquired twice by both observers. At the end of the collection period, the volumes were reviewed off-line by their respective observers to obtain the corresponding measurements: HC, AC, and FL and further referred to as datasets 3Dsenior and 3Djunior.

As part of the additional study, the volumes acquired by the junior observer were reset and processed again by the senior observer (dataset 3Dreview) to evaluate the specific impact of off-line processing on inter-observer reproducibility and to assess the impact of 3D acquisitions quality on the processing step. The intra-observer reproducibility was not studied as part of this additional study.

The two observers processed all the measurements in a random order to limit the impact of memorization effect and training bias on variability. These procedures produced four 2D-datasets for each fetus (two datasets 2Dsenior and two datasets 2Djunior) and five 3D-datasets (two datasets 3Dsenior, two datasets 3Djunior and one dataset 3Dreview). Examples of measurements performed by processing 3D acquisition (triplane view and post-reviewing view) and a flowchart of the data acquisition procedure are provided in Figs. 1 and 2, respectively.

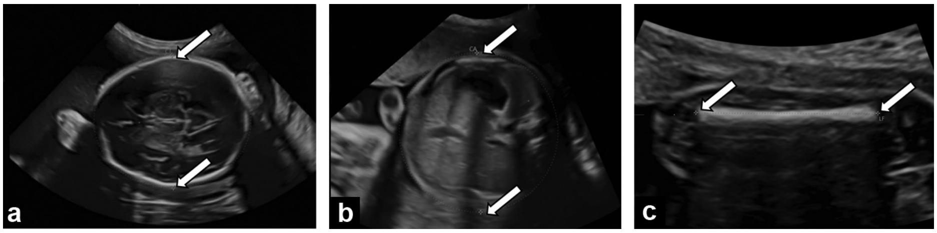

**Fig. 1.** Examples of measurements performed by three-dimensional processing acquisition (triplane view and post-reviewing view). a: head measurements; b: abdominal circumference; c: femur length. Arrows indicate caliper locations.

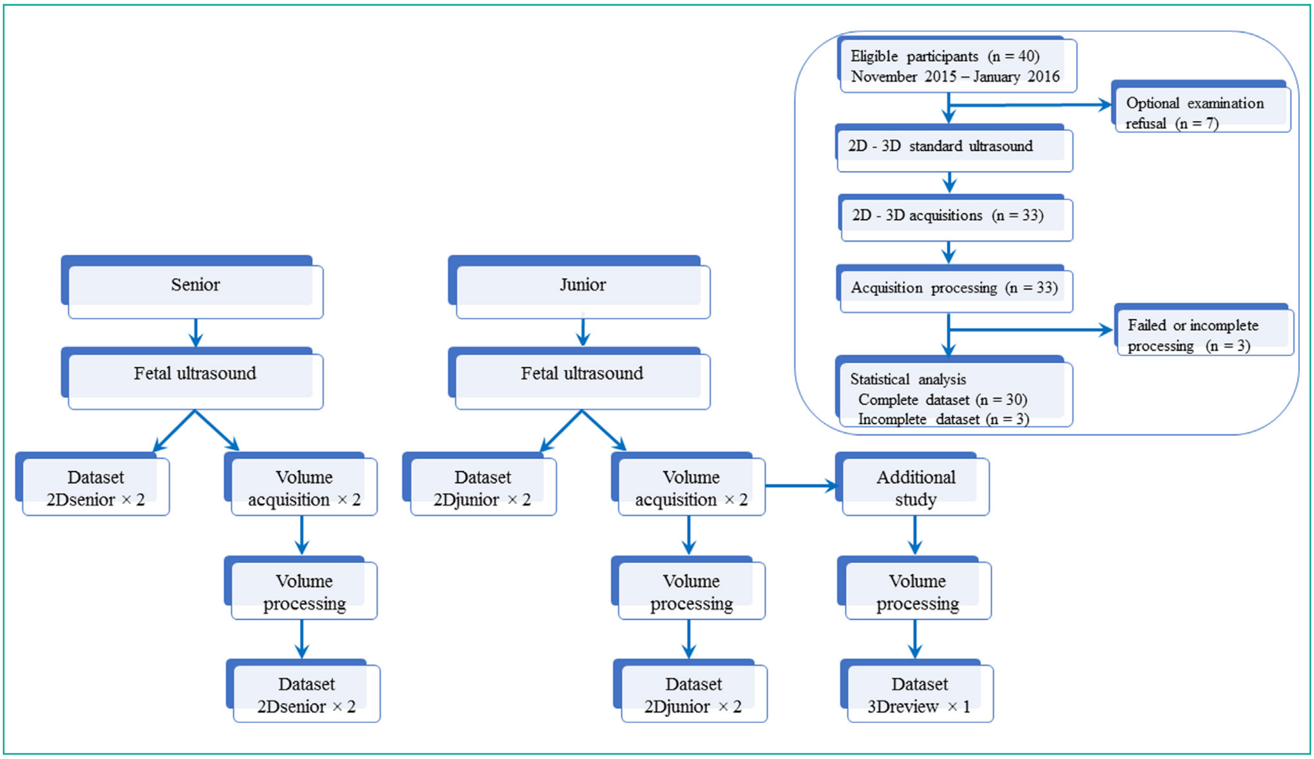

**Fig. 2.** Flowchart of the selection of patients and data acquisition procedures by the senior and junior observers.

For all women, indication for ultrasound examination, body mass index (BMI), maternal age, gestational age, complications or risk factors and fetal presentation were recorded.

### 2.4. Statistical analysis

The term "agreement" referred to the fact that two or more independent measurements of the same parameters were equal [14]. The term "repeatability" referred to the ability of a same observer to obtain similar results on iterative measurements performed on the same fetus [15]. The term "reproducibility" referred to the ability for different observers to obtain similar results measuring the same fetus. The terms "random error" and "systematic error' referred to the individual variability related to unpredictable changes due to changes in observer or technique and predictable changes related to a tendency to over- or under-estimate a magnitude.

Use of Z-scores to report fetal measurements resulted in variability remaining constant irrespective of fetal size [12]. All of the measurements were, therefore, converted into Z-scores using the references provided by the Intergrowth 21st study [1]. The following cut-offs, Z-score < –1.3 and Z-score > 1.3 (corresponding to measurements < 10th percentile and > 90th percentile) were used to assess the distribution of measurements in each series.

To underline the clinical impact resulting from a difference in fetal biometric measurements, the estimated fetal weight (EFW) was calculated using the Hadlock formula [16]. The Intergrowth 50° centile for 32 WG was used as the reference.

A minimum of 30 observations (*i.e.*, 30 fetuses) was required to provide a sample with a normal distribution as a part of a pilot study. A further 10% was added to mitigate the risk of missing data. Therefore, the protocol specified 33 fetuses needed to be scanned.

All quantitative data were expressed as mean ± standard deviation (SD) and ranges. Paired Welch two sample *t*-test was used for duration comparison. The differences between the different series of measurements were searched for with the Tukey's HSD test (honestly significant difference). Significance for difference was set at $P < 0.05$. The calculation of the intraclass correlation coefficients (ICC) with their 95% confidence interval (CI) allowed the assessment of the repeatability and reproducibility between the different series. Statistical analysis was performed using R software version 3.2.0 (R Core Team).

Agreement between the different series of measurements was evaluated using Bland–Altman plots [14].

## 3. Results

### 3.1. Patients

Of the initially 40 consecutive eligible women, measurements and acquisitions were performed on 33 women because 7 refused the additional examination. The mean maternal age was 28 years ± 3.9 (SD) years (range: 20–40 years) with a mean gestational age of 28 ± 5 (SD) WG (range: 20–40 WG). The mean body mass index (BMI) of the women was 27 ± 5.18 (SD) kg/m$^2$ (range: 17–36 kg/m$^2$) with 23 women having a BMI ≥ 25 kg/m$^2$.

Fetal measurements and volume acquisitions were performed during routine 2nd and 3rd trimester ultrasound appointments for 20 of the 33 women. The other 13 women were scanned to monitor fetal growth. The pregnancies progressed normally in 28/33 women (85%), with the other five women having the following complications or risk factors: gestational diabetes ($n = 3$), arterial hypertension ($n = 1$) and antiphospholipid antibodies syndrome ($n = 1$). Fetal presentation was cephalic in 20/33 women (61%), breech in 7/33 women (21%), and transverse in 6/33 women (18%). The flowchart of the study population recruitment is presented in Fig. 2.

Both observers acquired the necessary 2D images and 3D volumes for biometric measurements in all women. However, on processing his 3D volumes, the junior observer identified that on three occasions a volume of the shoulder and humerus had been mistakenly acquired instead of the pelvis and femur. The FL measurements corresponding to these three volumes were not calculated.

The mean time taken for fetal biometry in 2D & 3D is reported in Table 1. The senior observer obtained measurements significantly faster using both 2D images and 3D volumes than the junior observer. The use of 3D volume acquisition significantly reduced the "point of care" time for both observers. However, when the time for the complete procedure was analyzed (*i.e.*, 3D acquisition and processing) 3D procedure time was significantly longer than 2D procedure time for junior and senior observers ($P < 0.001$).

### 3.2. Distribution of measurements

There were no significant differences between the different observers' measurements of the HC and AC or between the 2D and 3D measurements (Table 2). The mean of the biometric measurements obtained by the junior observer using 2D ultrasound (dataset 2Djunior) was lower with a higher standard deviation than using 3D. And significant differences in 2D measurements were found for FL between the junior and the senior observer ($P = 0.038$).

The number of measurements < −1.3 Z-score or > 1.3 Z-score (below the 10th or above the 90th centile) was reduced for the junior observer when using 3D ultrasound. The concordance between the two observers for the number of measurements between < −1.3 and > 1.3 Z-score increased when the junior observer used 3D volumes.

### 3.3. Senior observer

The senior observer demonstrated high degrees of agreement between all measurement techniques, with a 95% CI of less than 1 Z-score for the three parameters (Fig. 3). This variability corresponded, in a clinically relevant manner, to 95% of EFWs being between 1534 g and 2039 g for a 1762 g fetus at 32 WG (Table 3). The repeatability of the measurements was high within all the series (ICC ≥ 0.90 for the datasets 2Dsenior and 3Dsenior) and an improvement was seen using 3D for the HC and FL parameters (Table 4).

### 3.4. Junior observer

The agreement between the reference measurements (dataset 2Dsenior) and the datasets 2Djunior and 3Djunior were weaker than those seen between the reference and the dataset 3Dsenior. For the HC and AC parameters, the 95% confidence interval was approaching 2 Z-scores in both 2D and 3D (Fig. 3). A systematic difference of approximately −0.4 and −0.2 Z-score with the reference measurements was found for the dataset 2Djunior for these

**Table 1**
Duration times for fetal biometry using two-dimensional and three-dimensional ultrasound by two observers.

| | 2D | 3D | | | *P*-value* |
|---|---|---|---|---|---|
| | Mean time 2D | Mean time volume acquisition | Mean time volume processing | Total mean time 3D procedure | |
| Junior observer | 6:02 ± 2:20<br>[1:55–11:50] | 2:49 ± 1:09<br>[1:30–7:21] | 8:50 ± 2:45<br>[4:47–15:55] | 11:40 ± 3:12<br>[6:54–19:51] | < 0.001 |
| Senior observer | 2:54 ± 0:44<br>[1:23–4:45] | 1:56 ± 0:32<br>[1:06–3:36] | 7:24 ± 1:28<br>[5:26–11:43] | 9:20 ± 1:40<br>[7:13–13:42] | < 0.001 |
| *P*-value* | < 0.001 | < 0.001 | 0.008 | < 0.001 | – |

Data are given in minutes:seconds and expressed as means ± standard deviations; numbers in brackets are ranges.

* Paired Welch two sample *t*-test was used for comparison.

**Table 2**
Distribution of measurements for 2Dsenior, 3Dsenior, 2Djunior and 3Djunior datasets.

| | Senior observer | | Junior observer | |
|---|---|---|---|---|
| | 2D | 3D | 2D | 3D |
| Head circumference | 0.212 ± 1.05<br>[−2.2–2.5]<br>3; 11[a] | 0.215 ± 0.94<br>[−2.1–2.9]<br>2; 5[a] | −0.134 ± 1.38<br>[−3.7–3.5]<br>14; 10[a] | 0.137 ± 1.19<br>[−2.2–3.1]<br>8; 8[a] |
| Abdominal circumference | 0.864 ± 1.15<br>[−1.7–3.5]<br>2; 22[a] | 0.93 ± 1.17<br>[−2.1–2.8]<br>3; 23[a] | 0.685 ± 1.16<br>[−1.8–3.5]<br>4; 25[a] | 0.755 ± 1.08<br>[−1.6–2.8]<br>3; 20[a] |
| Femoral length | 0.535 ± 1.23<br>[−2.9–3.6]<br>2; 13[a] | 0.472 ± 1.21<br>[−2.2–3.8]<br>2; 14[a] | −0.396 ± 1.65<br>[−6.6–3.8]<br>10; 2[a] | 0.006 ± 1.54<br>[−2.9–3.2]<br>5; 8[a] |

Data are expressed as means of Z-scores ± standard deviations; numbers in brackets are ranges.

[a] Numbers of measurements < −1.3 Z-score or 10 centiles and > 1.3 Z-score or 90 centiles for each dataset.

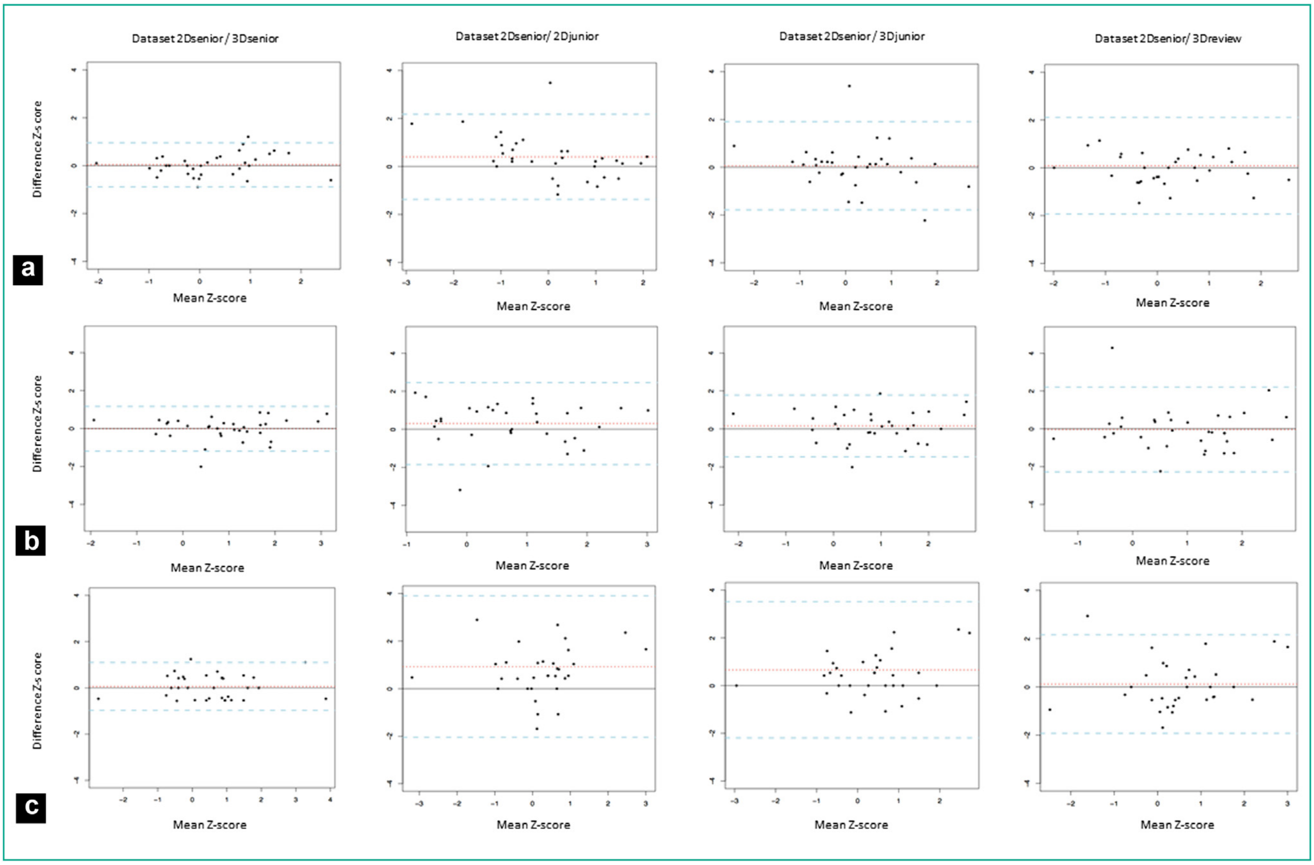

**Fig. 3.** Graphs show agreement with reference measurements (2Ddataset) for 3Dsenior, 2Djunior, 3Djunior & 3Dreview datasets. Head circumference (a), abdominal circumference (b) and femur length (c).

**Table 3**
Examples of fetal weight estimation (EFW) confidence interval at 32 weeks when comparing each technique to the reference set (Dataset 2Dsenior).

| Parameters | 50° percentile 32 WG | Dataset 2Dsenior versus 3Dsenior (95% CI)[a] | Dataset 2Dsenior versus 2Djunior (95% CI)[a] | Dataset 2Dsenior versus 3Djunior (95% CI)[a] | Dataset 2Dsenior versus 3Dreview (95% CI)[a] |
|---|---|---|---|---|---|
| Head circumference | 294.5 | 286.5–302.2<br>[−0.8–0.8] | 275–308.7<br>[−2–1.5] | 276.3–311.8<br>[−1.9–1.8] | 274.2–312.7<br>[−2.1–1.9] |
| Abdominal circumference | 273.9 | 259–289<br>[−1.1–1.2] | 240–301.3<br>[−2.5–2] | 249.5–294.3<br>[−1.8–1.5] | 245.5–304<br>[−2.1–2.2] |
| Femoral length | 59.4 | 56.8–61.8<br>[−1.1–1] | 49.9–64.2<br>[−4–2] | 51.1–64.4<br>[−3.5–2.1] | 54.5–64<br>[−2.1–1.9] |
| Fetal weight estimation[b] | 1762 | 1534–2039 | 1176–2281 | 1270–2215 | 1301–2341 |

Data are expressed in millimetres with corresponding Z-score in brackets. WG: weeks of gestation

[a] 95% confidence interval derived from Bland Altmann graphs in Fig. 3.

[b] Corresponding fetal weight estimation according to Hadlock 3.

**Table 4**
Repeatability & reproducibility of fetal measurements by to different observers using intraclass correlation coefficient.

| Repeatability[a] | Senior | | Junior | |
|---|---|---|---|---|
| | 2D | 3D | 2D | 3D |
| HC | 0.90 (0.81–0.95) | 0.93 (0.86–0.97) | 0.94 (0.88–0.97) | 0.94 (0.88–0.97) |
| CA | 0.95 (0.90–0.97) | 0.95 (0.87–0.97) | 0.88 (0.78–0.94) | 0.93 (0.87–0.97) |
| FL | 0.94 (0.89–0.97) | 0.97 (0.93–0.98) | 0.90 (0.81–0.95) | 0.96 (0.91–0.98) |
| Reproducibility[b] | Dataset 2Dsenior vs. 2Djunior | Dataset 2Dsenior vs. 3Djunior | Dataset 3Dsenior vs. 3Djunior | Dataset 2Dsenior vs. 3Dreview |
| HC | 0.75 (0.54–0.87) | 0.71 (0.49–0.85) | 0.79 (0.61–0.89) | 0.70 (0.46–0.84) |
| CA | 0.60 (0.34–0.78) | 0.78 (0.6–0.88) | 0.89 (0.79–0.95) | 0.60 (0.33–0.78) |
| LF | 0.45 (0.10–0.69) | 0.63 (0.36–0.8) | 0.63 (0.33–0.8) | 0.71 (0.48–0.85) |

Intraclass correlation coefficient was calculated from data given in Z-score and given along with their confidence intervals in parentheses. HC: head circumference; AC: abdominal circumference; FL; femur length.

[a] Comparison between repeated measurements of the same parameter by the same observer.

[b] Comparison between the first measurement of each dataset for each technique and observer.

two parameters, respectively. This difference was consistent with the average differences seen in Table 2 (underestimation of measurements within the 2Djunior dataset compared to the reference dataset).

Concerning the femur measurements, the agreement between the 2D and 3D series of the junior operator and the reference measurements were even weaker with a 95% CI > 3.5 Z-scores and an underestimation of measurements, compared to the reference dataset. This variability corresponds to 95% of EFWs being between 1176 g and 2281 g for a 1762 g fetus at 32 WG in 2D (Table 3). And, to 95% of EFWs being between 1270 g and 2215 g in 3D (Table 3). The repeatability of the measurements was raised in 3D with ICCs close to those seen within the 2Dsenior dataset (Table 4).

### 3.5. Reproducibility

The inter-operator intra-technique reproducibility was higher when both operators measured from 3D volumes rather than in 2D. A particular benefit was seen for AC and FL (ICC = 0.60 and 0.45 *vs.* 0.89 and 0.63 for 2D and 3D, respectively) (Table 4).

### 3.6. Impact of acquisition quality on 3D volumes processing

Analysis based on Bland Altmann plots showed that when the senior observer processed the 3D volumes acquired by the junior observer (dataset3Dreview) the agreement with the reference range (dataset 2Dsenior) was lower than the agreement of the 3Djunior dataset with the reference range. Similarly, reproducibility between 3Dreview & 2Dsenior was lower than 3Djunior & 2Dsenior for the measurements of HC and AC. Those results demonstrated a reduction in reproducibility in connection with the processing by the senior operator (Table 4). However, this same comparison for the FL shows a raised reproducibility within the 3Dreview dataset with a beneficial impact of the review by the expert operator (Table 4). The variability between 3Dreview & dataset 2Dsenior corresponds to 95% of EFWs being between 1301 g and 2341 g for a 1762 g fetus at 32 WG (Table 3).

## 4. Discussion

This pilot study demonstrated that, when taking fetal biometric measurements, agreement with the senior observer was higher when the junior observer used 3D ultrasound volumes. The distribution of measurements was closer to the reference range, and repeatability and reproducibility were increased. Taking fetal biometric measurements using 3D volumes was significantly slower for the junior observer when taking into account the entire procedure but allowed a significant shortening of the duration of the "point of care" examination.

On the opposite, using 2D ultrasound, the junior observer underestimated measurements. This was most notably seen for the FL. A possible explanation for this could be a confusion between identifying the humerus instead of the femur. This confusion was detected in three cases when reviewing the 3D volumes and these were excluded. As this potential confusion was not detected at the time of the 2D scan no measurements were excluded from the 2D dataset. In clinical practice this could lead to errors in obstetric management which could be mitigated by the use of 3D volumes.

In our study, there was a limited impact of the imaging technique (2D or 3D) on the biometric measurements obtained by the senior observer. The slight improvement of repeatability obtained using 3D for the measurements of HC and FL was balanced out by the duration of the procedure, which was significantly longer than the time taken to obtain the measurements in 2D.

The patients included in this study were done so in a consecutive manner without exclusion for growth abnormality. This explains the distribution seen within the cohort, with a large proportion of measurements < −1.3 Z-score and > 1.3 Z-score. This choice reflected a desire to study the impact of measurement techniques on a population representative of a standard clinic setting (physiological and pathological growth, variable fetal presentations, maternal BMI representative of the local population). Within this series, there were several overweight patients (23 patients out of 33 patients included), which is a factor that limits the quality of ultrasound measurements. However, it appears that the use of 3D ultrasound is feasible as volumes were obtained in all patients and indeed may have a beneficial impact when performing biometric measurements.

Regarding the differences in the range of measurements performed on 2D images and from 3D volumes, this has been already reported in a previous study and is potentially explained by the difficulties encountered in 2D measuring with the associated risk of random error and of a more significant under- or over-estimation of measurements (systematic errors) [9]. This bias seems to be limited by using the 3D technique especially for the junior observer [11], however, the specific composition of our cohort (a large proportion of measurements < −1.3 Z-score and > 1.3 Z-score) may have accentuated this effect. This finding calls for caution over the clinical interpretation of 3D measurements. Current reference ranges referring to 2D data are probably not superimposable on a 3D series.

The starting hypothesis of the additional 3D review study was to confirm that the senior observer's review of the junior observer's 3D volumes would result in increased reproducibility and agreement of the measurements. However, this point was not confirmed (inter-operator ICC was weaker for the HC and AC parameters for the 3Dreview dataset). The only advantage observed concerned the agreement of FL measurements with the reference measurements. These results underline that the experience of the observer appears to have a more significant impact when using 2D ultrasound than 3D ultrasound. This difference could be partly explained by the standardization of analyzing 3D volumes, limiting the observer-dependent nature of performing measurements. This would need to be confirmed on a more extensive series involving different observers from different institutions to reduce the interpretation bias related to each observer. However, these finding echoes the concerns recently put forward in the Beyond Ultrasound First Forum [5] regarding quality control of data and the evolution of ultrasound within different imaging techniques. It encourages research aimed at automating the measurement process, especially regarding the analysis of 3D volumes [17–19].

The use of Z-scores to compare ultrasound measurements prevents an over-estimation of the agreement and reproducibility for the measurements in early gestations (where a low variability is seen of measurements when presented in millimetres) and similarly an under-estimation of the same parameters at later gestations [12]. By using Z-scores in this study it allowed comparison of these data with previously published data. Variability seen between repeated measurements of the same parameter in 2D and 3D by the senior observer is equivalent to that seen in conventional 2D ultrasound of large cohorts. The use of Z-scores also allows a comparison of the 3D ultrasound measurements by the junior observer with data published in the same format, opening the possibility to subsequent comparisons to verify these preliminary results.

There are two main limitations to this study. The first relates to the small population size and the number of observers performing the scans. This limits the possibility of extrapolating the results to standard care. The second relates to the use of 2D measurements obtained by the senior observer as the gold standard for the reference range.

In conclusion, the results of this pilot study confirm prior results regarding the use of 3D ultrasound in reducing differences in fetal measurements due to the level of observer experience and also in reducing the duration of "point of care" examination for non-expert observers. These preliminary results support the possibility to standardize measurement procedures by using 3D ultrasound volumes, allowing us to reflect on the use of this technique in the training of physicians and sonographers, in the quality control of data and in the development of new measurement techniques.

## Human rights

The authors declare that the work described has been carried out in accordance with the Declaration of Helsinki of the World Medical Association revised in 2013 for experiments involving humans.

## Informed consent and patient details

The authors declare that this report does not contain any personal information that could lead to the identification of the patient(s) and/or volunteers.

## Funding

This work did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Author contributions

All authors attest that they meet the current International Committee of Medical Journal Editors (ICMJE) criteria for Authorship.

## CRediT authorship contribution statement

Conceptualization ideas: G.A., O.M., G.H., M.C.

Methodology development or design of methodology: G.A.,O.M., G.H., M.C.

Validation Verification: G.A., G.H., P.B.

Statistical analysis: G.H.

Investigation: G.A., M.C.

Data curation: G.A., M.C.

Writing–original draft: G.A., P.B., P.N.

Writing–review and editing: G.A., P.B., G.H., P.N., M.C., O.M.

Visualization: G.A., P.B., P.N.

Supervision: O.M., G.A.

Project administration: G.A, O.M, G.H.

Final draft approval: G.A., P.B., G.H., P.N., M.C., O.M.

## Disclosure of interest

The authors declare that they have no competing interest.

## References

[1] Papageorghiou AT, Ohuma EO, Altman DG, Todros T, Cheikh Ismail L, Lambert A, et al. International standards for fetal growth based on serial ultrasound measurements: the fetal growth longitudinal study of the INTERGROWTH-21st Project. Lancet 2014;384:869–79.

[2] Kha KS, Chien PFW. Evaluation of a clinical test. I: assessment of reliability. BJOG Int J Obstet Gynaecol 2001;108:562–7.

[3] 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 Gynecol 2015;46:14–20.

[4] Perni SC, Chervenak FA, Kalish RB, Magherini-Rothe S, Predanic M, Streltzoff J, et al. Intraobserver and interobserver reproducibility of fetal biometry. Ultrasound Obstet Gynecol 2004;24:654–8.

[5] Benacerraf BR, Minton KK, Benson CB, Bromley BS, Coley BD, Doubilet PM, et al. Proceedings: beyond ultrasound first forum on improving the quality of ultrasound imaging in obstetrics and gynecology. J Ultrasound Med 2018;37: 7–18.

[6] Sarris I, Ioannou C, Dighe M, Mitidieri A, Oberto M, Qingqing W, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. Ultrasound Obstet Gynecol 2011;38:681–7.

[7] Weerasinghe S, Mirghani H, Revel A, Abu-Zidan FM. Cumulative sum (CUSUM) analysis in the assessment of trainee competence in fetal biometry measurement. Ultrasound Obstet Gynecol 2006;28:199–203.

[8] 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;42:560–70.

[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;33:447–52.

[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;40:186–93.

[11] Yang F, Leung KY, Lee YP, Chan HY, Tang MHY. Fetal biometry by an inexperienced operator using two- and three-dimensional ultrasound. Ultrasound Obstet Gynecol 2010;35:566–71.

[12] Sarris I, Ioannou C, Chamberlain P, Ohuma E, Roseman F, Hoch L, et al. Intra- and interobserver variability in fetal ultrasound measurements. Ultrasound Obstet Gynecol 2012;39:266–73.

[13] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:5527.

[14] Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003;22:85–93.

[15] Soyer P. Agreement and observer variability. Diagn Interv Imaging 2018;99:53–4.

[16] 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;151:333–7.

[17] Hur H, Kim YH, Cho HY, Park YW, Won HS, Lee MY, et al. Feasibility of three-dimensional reconstruction and automated measurement of fetal long bones using 5D Long Bone. Obstet Gynecol Sci 2015;58:268–76.

[18] Ambroise Grandjean G, Hossu G, Bertholdt C, Noble P, Morel O, Grangé G. Artificial intelligence assistance for fetal head biometry: assessment of automated measurement software. Diagn Interv Imaging 2018;99:709–16.

[19] Ambroise Grandjean G, Hossu G, Banasiak C, Ciofolo-Veit C, Raynaud C, Rouet L, et al. Optimization of fetal biometry with 3D ultrasound and image recognition (EPICEA): protocol for a prospective cross-sectional study. BMJ Open 2019;9:e031777.