A comparison of ultrasound with magnetic resonance imaging in the assessment of fetal biometry and weight in the second trimester of pregnancy: An observer agreement and variability study

Jacqueline Matthew^{1,2}, Christina Malamateniou^{1,3}, Caroline L Knight^{1,4}, Kelly P Baruteau^{1,5}, Tara Fletcher^{1,6}, Alice Davidson¹, Laura McCabe¹, Dharmintra Pasupathy³ and Mary Rutherford¹

Abstract

Objective: To compare the intra and interobserver variability of ultrasound and magnetic resonance imaging in the assessment of common fetal biometry and estimated fetal weight in the second trimester.

Methods: Retrospective measurements on preselected image planes were performed independently by two pairs of observers for contemporaneous ultrasound and magnetic resonance imaging studies of the same fetus. Four common fetal measurements (biparietal diameter, head circumference, abdominal circumference and femur length) and an estimated fetal weight were analysed for 44 'low risk' cases. Comparisons included, intra-class correlation coefficients, systematic error in the mean differences and the random error.

Results: The ultrasound inter- and intraobserver agreements for ultrasound were good, except intraobserver abdominal circumference (intra-class correlation coefficient = 0.880, poor), significant increases in error was seen with larger abdominal circumference sizes. Magnetic resonance imaging produced good/excellent intraobserver agreement with higher intra-class correlation coefficients than ultrasound. Good interobserver agreement was found for both modalities except for the biparietal diameter (magnetic resonance imaging intra-class correlation coefficient = 0.942, moderate). Systematic errors between modalities were seen for the biparietal diameter, femur length and estimated fetal weight (mean percentage error = +2.5%, −5.4% and −8.7%, respectively, $p < 0.05$). Random error was above 5% for ultrasound intraobserver abdominal circumference, femur length and estimated fetal weight and magnetic resonance imaging interobserver biparietal diameter, abdominal circumference, femur length and estimated fetal weight (magnetic resonance imaging estimated fetal weight error >10%).

Conclusion: Ultrasound remains the modality of choice when estimating fetal weight, however with increasing application of fetal magnetic resonance imaging a method of assessing fetal weight is desirable. Both methods are subject to random error and operator dependence. Assessment of calliper placement variations may be an objective method detecting larger than expected errors in fetal measurements.

¹Centre for the Developing Brain, Division of Imaging Sciences and Biomedical Engineering, King's College London, King's Health Partners, St Thomas' Hospital, London, UK

²NIHR Biomedical Research Centre, Guy's and St Thomas' NHS Hospital Foundation Trust, London, UK

³Department of Family Care and Mental Health, Faculty of Education and Health, University of Greenwich, London, UK

⁴Department of Women and Children's Health, King's College London, King's Health Partners, St. Thomas' Hospital, London, UK

⁵Lysholm Department of Neuroradiology, National Hospital for Neurology and Neurosurgery, University College London Hospitals NHS Foundation Trust, London, UK

⁶Radiology Department, Guy's and St Thomas' Hospital NHS Foundation Trust, London, UK

Corresponding author:

Jacqueline Matthew, King's College London, St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK.
Email: jacqueline.matthew@kcl.ac.uk

Keywords

Biometry, fetal weight, fetus, observer variation, magnetic resonance imaging, ultrasonography, pregnancy trimester, second

Date received: 1 September 2017; accepted: 21 December 2017

Introduction

Accurate evaluation of fetal size and growth is essential for the delivery of good quality antenatal care, and ultrasound (US) measurements play a central role. When an US scan indicates that a fetus is appropriately grown, this suggests good intrauterine health. Additionally, accurate antenatal detection of a growth abnormality may raise suspicions of a variety of fetal and maternal conditions which include pre-eclampsia, fetal growth restriction, gestational diabetes, macrosomia, infection and syndromic or genetic conditions.^{1,2} The information about fetal size may act as a threshold for clinicians to offer further investigations such as Doppler US, blood tests, amniocentesis or be used to plan the timing of delivery.³ However, US is known for its large random errors in fetal measurement and low sensitivity for detecting growth disturbances.^{2,4} Furthermore, there is growing evidence that magnetic resonance imaging (MRI) can result in estimated fetal weight (EFW) with far less error than US, particularly when using volumetric methods.⁵⁻⁷ Few studies have assessed the validity of MRI by radiologists for the

measurement of fetal biometry compared to US by sonographers.⁸⁻¹⁰ Additionally, a literature search did not reveal studies that had performed a comprehensive variability and method comparison of US and MRI for fetal biometry and EFW. It is also noted that reporting standards of method comparison studies vary widely which limits their interpretation.¹¹⁻¹⁴

Fetal MRI is a highly specialised modality for fetal diagnosis and is well established for fetal central nervous system (CNS) anomalies. A systematic review of 13 peer-reviewed articles, found that MRI provided supplementary information to US and resulted in a change in clinical management in 30% of cases – referral indications were numerous.^{15,16} However, MRI is also increasing in its remit for fetal evaluation of anomalies outside the CNS e.g. diaphragmatic hernia or pulmonary anomalies, particularly when US is limited by reduced amniotic fluid, maternal obesity or in the presence of equivocal US findings.¹⁶⁻¹⁹ A survey conducted by the International Society of Ultrasound in Obstetrics and Gynaecology (ISUOG), found that at least one to two centres in 27 countries were performing fetal MRI with the quality of

Figure 1. US head circumference (HC) plane.

Figure 2. MRI slice-to-volume reconstruction (SVR) head circumference (HC) plane.

Figure 3. US biparietal diameter (BPD) plane.

Figure 6. MRI abdominal circumference (AC) plane.

Figure 4. MRI slice-to-volume reconstruction (SVR) biparietal diameter (BPD) plane.

Figure 7. US femur length (FL) plane.

Figure 5. US abdominal circumference (AC) plane.

Figure 8. MRI femur length (FL) plane.

imaging sequences used and operator experience varying widely. In the UK, fetal MRI is offered by few local tertiary units (currently approximately six UK wide), and may involve outsourcing of image reporting to experienced specialists. ISUOG also suggests that a standardised and complete assessment of fetal anatomy is feasible with MRI, however, its current remit is to complement an expert US examination.¹⁶

As the use of clinical fetal MRI increases, an assessment of fetal/biometry weight is desirable but under tested across gestational ages (GAs). Previous studies of EFW have almost exclusively focussed on fetal MRI late in gestation, however women may be referred for a fetal MRI scan soon after the 20-week anomaly US scan for further assessment.^{3,20} The aim of this study is to compare the intra and interobserver variability of

Table 1. Difference in the mean ultrasound (US) and magnetic resonance imaging (MRI) biometric measurements and estimated fetal weight (EFW).

Measurement	<i>n</i>	US, Mean, mm	MRI _{biom} , Mean, mm	Absolute difference, mm (95% confidence interval, CI)	Percentage difference, %	Paired <i>t</i> -test (<i>p</i> value)
Biparietal diameter, BPD	30	58.8	60.2	−1.5 [−2.2 to −0.8]	−2.5	<0.001
Head circumference, HC	30	215.5	215.4	0.6 [−1.4 to 1.5]	0.3	0.9
Abdominal circumference, AC	42	191.4	190.3	1.1 [−2.3 to 4.5]	0.6	0.5
Femur length, FL	33	42.0	39.7	2.2 [1.0 to 3.7]	5.4	0.001
Estimated fetal weight, EFW, g	25	647.1	593.3	53.8 [19 to 89]	8.7	<0.05

Table 2. Intraobserver and interobserver agreement. Intraclass Correlation Coefficient (ICC) with 95% confidence interval (CI).

Fetal measurement (<i>n</i>)	Ultrasound (US) ICC (95% CI)	MRI ICC (95% CI)
Intraobserver		
BPD (28)	0.982, good, [0.959–0.992]	0.995, excellent [0.988–0.997]
HC (28)	0.952, good, [0.580–0.986]	0.997, excellent [0.994–0.999]
AC (40)	0.880, poor, [0.772–0.937] <i>Significant proportional bias, $p < 0.05$</i>	0.994, excellent [0.988–0.997]
FL (31)	0.978, good [0.944–0.990]	0.989, good [0.975–0.995]
EFW (23)	0.972, good [0.547–0.993]	0.983, good [0.961–0.993]
Interobserver		
BPD (28)	0.974, good [0.808–0.992]	0.942, moderate [0.860–0.974]
HC (28)	0.971, good [0.938–0.987]	0.983, good [0.963–0.992]
AC (40)	0.967, good [0.896–0.982]	0.973, good [0.949–0.986]
FL (31)	0.990, good [0.979–0.995]	0.978, good [0.955–0.990]
EFW (23)	0.988, good [0.965–0.995]	0.964, good [0.905–0.986]

BPD: biparietal diameter; HC: head circumference; AC: abdominal circumference; FL: femur length; EFW: estimated fetal weight.

Figure 9. Intraobserver biparietal diameter (BPD) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

US and MRI in the assessment of common fetal biometry and EFW in the second trimester.

Design and methods

The intelligent fetal imaging and diagnosis project (iFIND) is a large scale, single centre observational imaging and engineering project, whose aim is to use novel technologies to improve diagnosis and detection rates in the second trimester of pregnancy.

The study is divided into iFIND-1 where 10,000 clinical mid-trimester anomaly US scans are recorded for the purposes of machine learning and big data analysis. The second part of the study is iFIND-2 where a smaller subset of participants are scanned, and includes a dedicated 2D and 3D US, as well as a MRI research scan on each fetus. The iFIND-2 paired data sets are obtained within 0–3 days. The images were retrospectively and consecutively collected from the iFIND-2 data sets with a normal anomaly scan result. The image planes pre-selected included, the biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC) and femur length (FL) (see Figures 1 to 8 for image planes and measurement criteria examples). To

calculate the EFW for each fetus, the Hadlock D formula, including the HC, AC and FL measurements, were used as recommended by the British Medical Ultrasound Society and ISUOG.^{20–22} Whilst the BPD measurement is useful to assess head shape, its variability in measurement suggests it should not be used in routine EFW calculation, however there is debate in the literature about the best formula to use.²³

The US system was a Philips Epiq (Philips Healthcare, Best, Netherlands) and the participants were examined by one of two operators (JM or CK), a CASE accredited sonographer with 10 years' experience and an obstetrician with six years' UK scanning experience respectively. A 6–1 mHz matrix probe was used to scan all patients. The MRI scanner used for all participants was a Philips Ingenia 1.5 Tesla system (Philips Healthcare, Best, Netherlands). Motion-corrected MRI slice to volume reconstructions (SVRs) of the fetal head were used to find a transventricular plane comparable to US imaging.²⁴ An US and an MRI database of anonymised paired scans were compiled using the Osirix image review software for offline/remote review (version 7.5, Geneva, Switzerland). The databases were duplicated then the images reordered

Figure 10. Interobserver biparietal diameter differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

randomly, ready for a repeat review by the observers after 2.5 weeks with the aim of reducing any recall bias. All reviewers were provided with face-to-face training and guidance notes about; which views to record; the use of the Osirix and optimal viewing conditions for the offline review.

Using both of the US databases, one sonographer (TF, a UK trained sonographer with three years scanning experience) performed repeated measures (blinded to MRI and any previous measurements), this was used for US intraobserver (within) calculations. An obstetrician (CK) independently performed one US reading from the first database, for interobserver (between) calculations. Using both the MRI databases one radiologist (KP, five years fetal MRI clinical experience) performed repeated measures (blinded to the US and any previous measures) and a fetal imaging research radiographer (CM, 10 years fetal MRI research experience) independently performed one MRI reading from the first database. The observers also recorded a three-scale image quality score for each image (1 = poor, 2 = satisfactory and 3 = good). Data were collected on an Excel spreadsheet and all supplementary materials and raw data were deposited in a University Research Data Management System.

Image plane selection and calliper placement criteria were obtained from the NHS fetal anomaly screen programme guideline.²⁰

In the transventricular view (Figures 1 and 2), the image plane was at the level of the cavum septum pellucidum anteriorly (*) and the lateral ventricular horn posterior containing the choroid plexus (^). The falx cerebri was mid-line (") and the head an oval shape. The ellipse tool was used to measure around the outer table of the skull, being careful not to include any subcutaneous fat. The MRI transventricular view was carefully selected from SVR²⁴ obtained from T2 dynamic sequences (TR/TE = longest/80, slice Th/gap = volume/−1.25) which were manipulated in Osirix using the multiplanar reconstruction (MPR) mode.

In the same image plane as the for the HC measurement, the BPD was measured from the outer table of the skull to the outer table of the skull at the widest part for both MRI and US (Figures 3 and 4).

The AC measurement (Figures 5 and 6) was obtained with an ellipse tracing. The image plane was at a level including the part of the fetal liver (*), the fetal stomach (^), the portal sinus of the umbilical vein ("), three bony points of a vertebra in cross section (+), a circular abdominal

Figure 11. Intraobserver head circumference (HC) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (+/- 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

appearance, circular aorta (>) and with a short length of a rib, i.e. ‘unbroken’ (‘). The MRI sequence most commonly selected with the correct plane, was a T2 fast spin echo sequence of the transverse uterus (TR/TE = 920/90, slice Th/gap = 4/0).

The FL (Figures 7 and 8) was measured by placing the callipers at the end of the diaphysis in a view where the femur does not appear foreshortened (solid line). Care was taken to avoid measuring the cartilaginous epiphysis at either end of the femur and also to avoid the greater trochanter which otherwise would falsely elongate the measurement. The MRI sequence most commonly found to have a clear view of the femur in the correct plane was a diffusion weighted imaging (DWI) sequence in the B0 field i.e. before the diffusion weighting was applied, (TR/TE = 4000/89, slice Th/gap = 5/0). Some MRI femur views were well visualised using a gradient echo echoplanar imaging sequence.

Statistical analysis

The data were analysed using the statistical packages, SPSS (version 23, SPSS Inc, Chicago, Ill, USA) and

Excel (version 14.4.7, Microsoft Corp. Redmond, Washington, USA). The EFW was calculated using the Hadlock formula D.²⁵ A power calculation determined that a sample size of 31 was required to give a power of 80% for an error of 5% to detect an effect size of 1 mm difference (assuming a standard deviation (SD) of 8 mm). Normality testing was performed to ensure assumptions were met for statistical analysis and to identify any obvious outliers.

To assess systematic error between the modalities, the mean difference in measurement from the two observers per modality was compared for each parameter (BPD, HC, AC, FL and EFW). A two-tailed paired *t*-test was performed to compare the means.

To test the intra- and interobserver agreement, the average measures intra-class correlation coefficient (ICC) was used. Suggested cut off limits proposed in the literature for fetal studies guided interpretation.²⁶

Bland Altman plots were used to graphically assess the mean difference and the limits of agreement, LoA. A linear regression coefficient was used to determine if there was a statistically significant proportional bias in the error as the size increased.

Figure 12. Interobserver head circumference (HC) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

Random error was compared between modalities using the LoA (± 1.96 SD of the mean) as a marker of intra and interobserver variability and a two-tailed paired t -test was performed.

Finally, to allow the clinical significance to be interpreted more readily, the proportion of cases falling outside of a calliper placement error threshold was calculated. Arbitrary thresholds were determined by previous examples of expected error in the literature.⁴ In addition, a SD threshold for each parameter was determined using 1 SD of the US intraobserver measurements observed. A number and percentage of cases falling outside of the threshold ranges were tabulated and compared between MRI and US.

Results

Fifty-three consecutive iFIND-2 participants were recruited between November 2015 and April 2016 and had their fetal imaging studies reviewed for inclusion. Forty-four participants (83%) had fully paired data sets, and of these 25 (47%) had complete datasets and 19 (36%) were partially complete. Nine cases were excluded from the study because: four did not attend both scans; two had no

transventricular US scan plane available; two had failed or poor quality MRI head SVRs and two had missing US images.

The GA was a mean of 23.5 weeks (range 20.3–25.7). The body mass index (BMI) was a mean of 26.3 kg/cm (range 22.2–38.4 kg/cm), with three cases above 30 kg/cm (clinically obese). Sixty-eight percent of US and MRI scans were on the same day, 4% had a two-day interval and 24% had a three-day interval. Eighty-four percent of the US scans had a satisfactory mean image score and 16% had a good score. For MRI, 8% had a poor mean score, 80% had a satisfactory score and 12% had a good score.

Table 1 demonstrates that MRI systematically measured the BPD larger than US (mean percentage error = 2.5%, or 1.5 mm, $p = 0.001$), and the FL smaller than US (mean percentage error = -5.4%, or -2.2 mm, $p = 0.001$). MRI systematically measured the EFW smaller than US (mean percentage error = -8.7%, or -53.8 g, $p < 0.05$). The mean measurements of the HC and AC compared well between modalities.

After normality testing, two outliers were removed from the data set for the subsequent analysis. One was an obvious data input error for the MRI BPD (case 6) and one was a significant measurement error due poor

Figure 13. Intraobserver abdominal circumference (AC) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (+/- 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

image quality of a T2 sequence for bone (case 18). Only one other outlier was identified for US AC, however it was unclear if this was a data input error or a true observer measurement so was kept for the remaining analysis (case 41).

Table 2 shows that MRI had excellent intraobserver agreement for BPD, HC and AC and good FL agreement, with all ICC results scoring higher than US. The intraobserver HC and AC had non-overlapping confidence intervals between modalities suggesting significant differences in agreement. US had good intraobserver agreement for all parameters except AC which scored poorly (ICC = 0.880). In addition, there was significantly less agreement for the US AC intraobserver measurement as the AC absolute size increased ($p < 0.05$).

For interobserver agreement, US and MRI both had good agreement for all parameters except for the MRI BPD (moderate ICC = 0.942), however all parameters had overlapping 95% confidence intervals, suggesting no significant difference in agreement.

The Bland Altman plots in Figures 9 to 18 show the absolute difference in millimetres between two measurements for each individual case. The MRI and US differences are overlaid on the same plot with a central

mean difference line and a LoA line above and below to represent 95% of the variance. Only intraobserver AC showed an increase in variation with size, with a marginal increase seen with intraobserver FL and EFW that was not significant. The LoA varied between parameters, with a tendency for MRI LoA to be narrower than US for intraobserver measures and wider for interobserver measures.

In Table 3, the LoA (random error) are explored further, demonstrating statistically significant differences for the intraobserver LoA for HC and FL, with MRI having less variation than US ($p < 0.05$). There were significant differences in the interobserver LoA for AC and FL, with MRI having more variation than US ($p < 0.05$). Parameters where the mean variation was above an arbitrary 5% percentage error threshold, included the intraobserver US measures of AC, FL and EFW (8.7%, 5.0% and 6.5%, respectively) and MRI EFW (13.6%). For interobserver measures, the parameters for MRI with a mean percentage error above 5% include BPD, AC, FL and EFW (5.0%, 5.5%, 6.9% and 8.9% respectively). For US, only interobserver EFW had a mean percentage error of more than 5% (6.3%).

Figure 14. Interobserver abdominal circumference (AC) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

Table 4 demonstrates that for intraobserver measures, more US cases that fell outside of the anticipated error range when compared to MRI (33 and 4 cases, respectively). For interobserver error, 15 MRI cases and 3 US cases fell outside the expected threshold for error.

Table 5, with narrower thresholds (based on intraobserver US SDs), demonstrated MRI measurements that had less cases falling out of range compared to US for intraobserver measures (62 US cases vs. 33 MRI cases). For interobserver cases, there were 56 US cases and 55 MRI cases in total with larger random error. For intraobserver and interobserver EFW with SD thresholds, the MRI measurements appeared to have a slightly greater random error than US with more cases outside of the 5% threshold (10 vs. 8, and 6 vs. 4, respectively).

Discussion

This study sought to comprehensively compare the intra- and interobserver variability between MRI and US for fetal measurements and EFW. The calliper placement errors for both US and MRI were in many cases above 5%, however these random errors observed

were expected to be smaller than in clinical practice because of the highly controlled conditions (one pre-selected image plane used retrospectively). Both modalities had cases falling outside of previously published error thresholds for fetal measurements.⁴ The causes of random errors in the US measurements are multifactorial in origin and include fetal position, maternal adiposity, sonographer experience, equipment specification and reduced amniotic fluid which could limit the view.^{4,27,28} Observer variation, is known to have a major impact on the precision of US fetal measurements, with electronic calliper placement on an image accounting for 58–80% of the error, and having more impact than maternal adiposity or fetal position.^{4,5} This highlights the need for thorough operator training and audit but also the need for technological development of more quantifiable and less subjective assessments, for example the use of z-scores similar to those used in first trimester nuchal translucency measurements.^{29,30}

Sarris et al. in 2012, investigated fetal biometry variation in 175 cases with three experienced sonographer observers, and found intraobserver variation to be consistently smaller than interobserver variation.⁴ The poor US intraobserver random error in this study was

Figure 15. Intraobserver femur length (FL) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

surprising and is an example of operator dependence that could have clinical implications especially when serial scans are performed, often by different operators. Operator experience has been demonstrated to have only a small impact on variation and therefore standardisation exercises before taking fetal measurements have been suggested.³¹ For MRI, the wider interobserver error was expected as these fetal measurements are rarely measured routinely and the operator experience thus more limited. There is a case for objective validation of measurement landmarks for training purposes and also for US and MRI specialists to work across disciplines, developing practice that compliment one another. The systematic errors observed in the study suggest modality-specific growth charts for MRI are needed, however currently there are none universally agreed or validated for clinical use. This is largely because MRI is a relatively new tool with less reference data available; most fetal MRI examinations are for the brain or spine where the technique is better established, and; there is an assumption that the routinely utilised US reference data and growth charts are suitable to use across the two modalities.^{9,32}

The EFW variability suggests that the random errors in fetal measurements will often compound the systematic errors of the mathematical equation, whether using US or MRI.³³ This is especially true for AC measurements on which the EFW formulae heavily rely. Indeed, Khel et al., 2012 suggests that the current accuracy of EFW has reached its limits, and that novel approaches to US technology must be considered to reduce clinical errors. 3D US fractional limb volumes have been used as an additional parameter for EFW, however as yet, there is a paucity of diagnostic accuracy tests to validate its use clinically,^{27,34–36} and reductions in post-processing time are needed to make this a useful tool in the future.^{1,2} Significant variation in EFW calculations has clinical implications because currently US is not recommended to screen the low-risk population for growth disturbances due to poor sensitivity and specificity.³⁷ Additionally, errors in the formula occur at the extremes of the weight range, due to changes in the soft tissue fat/muscle ratio of a compromised fetus, and may result in an overestimation of weight in small fetuses and an underestimation of weight in large fetuses when accurate depiction is most clinically important.³⁸

Figure 16. Interobserver femur length (FL) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

There is growing evidence that volumetric MRI can result in EFWs compared to birthweight with less random error than US, reported as low as 1–3% versus up to 7% for US.^{5,9,39,40} Moreover, MRI can negate some of US's technical drawbacks because maternal size, amniotic fluid and fetal position are less of a problem due to MRI's increased field of view. Still, fetal movements in MRI can cause image degradation, particularly in the second trimester when the fetus is more active, and result in a poorer signal-to-noise ratio. However, MRI has superior soft tissue contrast and improved boundary definition when placing electronic callipers for measurement or when outlining segments of the fetal body to calculate a volume.

The use of MRI is limited by its expense, lack of expertise and scanner availability, as well as the small evidence base of MRI's advantages over obstetric US for non-CNS anomalies. Here, differences in the imaging physics of each method most likely account for the systematic error in the mean measurement between modalities.^{9,11} For example, the use of T2 weighted MRI images could mean the anatomical landmarks are slightly different to US, e.g. more subcutaneous scalp tissue may have been included in the BPD view

due to the poorer bone definition. Distortion effects of the echoplanar imaging sequences used to select a FL plane on MRI would have resulted in the smaller and more variable FL measure. Technical refinement of MRI sequences is necessary for a comparable and representative assessment of fetal anatomy.

A major strength of the study was the use of recommended reporting guidance and statistics, thus avoiding some of the heterogeneous methods used in previous publications.^{11–13,26,41} As a retrospective study, there were limitations in the sample size and image sequences or views obtained for review, therefore the statistics should be interpreted with caution. Also, a prospective study would mean real-time US (as in clinical practice) could reveal the true variability. Furthermore, US was used as the reference standard to compare MRI – but it is well documented that the technique is prone to large random errors. Also, due to the small numbers, no statistical assessment of confounders (e.g. BMI, image quality or fetal position) could be attempted.

Future research should investigate the role of whole fetal body volume segmentation by MRI (or US) in the assessment of fetal weight as the technology continues to develop at a rapid pace.^{5,27,34} Methods to assess

Figure 17. Intraobserver estimated fetal weight (EFW) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

Figure 18. Interobserver estimated fetal weight (EFW) differences.

Figures 9–18 are Bland Altman Plots of US compared to MRI, showing mean absolute error, in mm, and limits of agreement, LoA, (± 1.96 SD) above and below the mean. US = blue circles, $\circ$ and solid lines—, MRI = green crosses, $\times$ and dashed line ---

Table 3. Differences in random error between ultrasound (US) and magnetic resonance imaging (MRI) fetal measurements and biometry-derived estimated fetal weight (EFW) (paired *t*-test).

Fetal measurement (<i>n</i>)	Intraobserver			Interobserver		
	US	MRI	<i>P</i> value	US	MRI	<i>P</i> value
Absolute error, mm (± 1.96 SD)						
BPD (28)	1.4	1.1	0.3	1.3	3.1	0.8
HC (28)	7.1	2.5	<0.05	7.9	6.2	0.6
AC (40)	18.0	5.1	0.09	8.7	10.3	<0.05
FL (31)	2.1	1.6	<0.05	1.6	2.6	0.99
EFW, g (23)	45.6	73.2	0.1	43.0	54.2	0.6
Percentage error, % (± 1.96 SD)						
BPD (28)	2.4%	1.8%	0.3	2.2%	5.0%	0.9
HC (28)	3.2%	1.2%	<0.05	3.6%	2.8%	0.6
AC (40)	8.7%	2.7%	0.1	4.6%	5.5%	<0.05
FL (31)	5.0%	4.2%	<0.05	3.8%	6.9%	0.97
EFW, g (23)	6.5%	13.6%	0.2	6.3%	8.9%	0.8

BPD: biparietal diameter; HC: head circumference; AC: abdominal circumference; FL: femur length; EFW: estimated fetal weight; SD: standard deviation.

Table 4. Differences in proportion of ultrasound (US) and magnetic resonance imaging (MRI) cases falling outside of arbitrary error threshold.

Arbitrary cut off			Intraobserver > threshold		Interobserver > threshold	
Parameter	Total cases	Threshold values. Intra/inter (mm)	US = <i>n</i> (%)	MRI <i>n</i> (%)	US <i>n</i> (%)	MRI <i>n</i> (%)
BPD	28	1.4/2.2	1 (4)	1 (4)	0 (0)	6 (21)
HC	28	5.2/8.0	11 (39)	0 (0)	2 (7)	1 (4)
AC	40	7.9/11.0	16 (15)	0 (0)	1 (3)	3 (8)
FL	31	2.1/2.5	3 (10)	0 (0)	0 (0)	2 (6)
EFW	23	66.1 g (10%)	2 (13)	3 (0)	0 (0)	0 (0)
Total number of cases measures of range (<i>n</i> = 150)			33 (22)	4 (3)	3 (2)	12 (8)

BPD: biparietal diameter; HC: head circumference; AC: abdominal circumference; FL: femur length; EFW: estimated fetal weight.

Table 5. Differences in proportion of ultrasound (US) and magnetic resonance imaging (MRI) cases falling outside of 1 standard deviation (SD) error threshold.

1 SD cut off			Intraobserver > threshold		Interobserver > threshold	
Parameter	Total cases	Threshold values (mm)	US = <i>n</i> (%)	MRI <i>n</i> (%)	US <i>n</i> (%)	MRI <i>n</i> (%)
BPD	28	0.6	13 (46)	8 (29)	16 (57)	17 (61)
HC	28	2.8	15 (54)	1 (4)	10 (36)	6 (21)
AC	40	3.3	19 (48)	8 (20)	21 (53)	17 (43)
FL	31	1.1	7 (23)	6 (19)	5 (16)	9 (29)
EFW	23	33.1 g (5%)	8 (35)	10 (43)	4 (17)	6 (26)
Total number of measures out of range (<i>n</i> = 150)			62 (41)	33 (22)	56 (37)	55 (37)

BPD: biparietal diameter; HC: head circumference; AC: abdominal circumference; FL: femur length; EFW: estimated fetal weight.

measurement variability as part of individual and departmental audit should also be investigated as part of audit or training programmes, with the aim of providing much needed objective quality assurance.

Conclusion

US remains the modality of choice when assessing biometry and estimating fetal weight. However with increasing applications of fetal MRI, a method of assessing fetal growth and weight is desirable. Both methods are subject to random error and operator dependence, with US being more operator dependent and MRI being an immature modality for common biometry. Since, EFW is affected by the variability of 2D measures, novel approaches, such as 3D volumetric methods in MRI, need further investigation if clinical errors are to be reduced in the future. The assessment of calliper placement variations, may be an objective method of detecting larger than expected errors in fetal measurements.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) declared receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Wellcome EPSRC Centre for Medical Engineering at King's College London (WT 203148/Z/16/Z) and by the Wellcome Trust IEH Award [102431]. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and

St Thomas' NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Ethical approval

The project has been granted NHS Research and Development approval and ethics approval, National Research Ethics Service reference number = 14/LO/1806 (trial registry numbers: UKCRN ID = 18283, ISRCTN = 16542843). All participants gave written and informed consent.

Guarantor

JM

Contributors

JM proposed the research question, researched the literature for the study and wrote the first draft of the manuscript. JM designed the study with guidance from MR, CM and DP who also contributed to the intellectual content and final version of the manuscript. JM, CLK, TF, KPB and CM contributed to data acquisition and analysis. AD and LM provided MRI reconstructions for the purposes of the research. JM performed the statistical analysis. All authors approved the final version of the manuscript.

Acknowledgements

Many thanks to the radiographers and sonographers at St Thomas' Hospital who contributed to the MRI and US data acquisition. Also, thank you to Trevor Murrells from the department of Nursing and Midwifery, KCL, for his expertise and support in the design of the statistical analysis. This manuscript formed part of a Masters in Clinical Research at King's College London.

References

- Malin GL, Bugg GJ, Takwoingi Y, et al. Antenatal magnetic resonance imaging versus ultrasound for predicting neonatal macrosomia: a systematic review and meta-analysis. *Int J Obstetri Gynaecol* 2016; 123: 77–88.
- Dudley N. A review of ultrasound fetal weight estimation in the early prediction of low birthweight. *Ultrasound* 2013; 21: 181–186.
- RCOG. *Termination of pregnancy for fetal abnormality. A working party report*. London, UK: Royal College of Obstetricians and Gynaecologists, 2010.
- Sarris I, Ioannou C, Chamberlain P, et al. Intra- and interobserver variability in fetal ultrasound measurements. *Ultrasound Obstet Gynecol* 2012; 39: 266–273.
- Kacem Y, Cannie MM, Kadji C, et al. Fetal weight estimation: comparison of two-dimensional US and MR imaging assessments. *Radiology* 2013; 267: 902–910.
- Zaretsky MV, Reichel TF, McIntire DD, et al. Comparison of magnetic resonance imaging to ultrasound in the estimation of birth weight at term. *Obstet Gynecol* 2003; 189: 1017–1020.
- Baker PN, Johnson IR, Gowland PA, et al. Fetal weight estimation by echo-planar magnetic resonance imaging. *Lancet* 1994; 343: 644–645.
- James JR, Khan MA, Joyner DA, et al. MR biomarkers of gestational age in the human fetus. *Magnatom Flash* 2012; 1: 112–118.
- Parkar AP, Olsen OE, Gjelland K, et al. Common fetal measurements: a comparison between ultrasound and magnetic resonance imaging. *Acta Radiol* 2010; 51: 85–91.
- Hatab MR, Zaretsky MV, Alexander JM, et al. Comparison of fetal biometric values with sonographic and 3D reconstruction MRI in term gestations. *AJR Am J Roentgenol* 2008; 191: 340–345.
- Bartlett JW and Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol* 2008; 31: 466–475.
- Coelho Neto MA, Roncato P, Nastri CO, et al. True Reproducibility of UltraSound Techniques (TRUST): systematic review of reliability studies in obstetrics and gynecology. *Ultrasound Obstet Gynecol* 2015; 46: 14–20.
- Bland JM and Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol* 2003; 22: 85–93.
- Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307–310.
- Rossi AC and Prefumo F. Additional value of fetal magnetic resonance imaging in the prenatal diagnosis of central nervous system anomalies: a systematic review of the literature. *Ultrasound Obstet Gynecol* 2014; 44: 388–393.
- Prayer D, Malinger G, Brugger PC, et al. ISUOG practice guidelines: performance of fetal magnetic resonance imaging. *Ultrasound Obstet Gynecol* 2017; 49: 671–680.
- Gonçalves LF, Lee W, Mody S, et al. Diagnostic accuracy of ultrasonography and magnetic resonance imaging for the detection of fetal anomalies: a blinded case-control study. *Ultrasound Obstet Gynecol* 2016; 48(2): 185–192.
- Pugash D. Fetal MRI: the sonographer's view. *Top Magn Reson Imaging* 2011; 22: 91–99.
- Gholipour A, Estroff JA, Barnewolt CE, et al. Fetal MRI: a technical update with educational aspirations. *Concepts Magn Reson Part A Bridging Educ Res* 2014; 43: 237–266.
- PHE. Fetal anomaly screening programme standards 2015-16. London, UK: Public Health England, 2015.
- Hadlock FP, Harrist RB, Sharman RS, et al. Estimation of fetal weight with the use of head, body, and femur measurements – a prospective study. *Am J Obstet Gynecol* 1985; 151: 333–337.
- Salomon LJ, Alfrevic Z, Berghella V, et al. Practice guidelines for performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound Obstet Gynecol* 2011; 37: 116–126.
- Loughna P, Chitty L, Evans T, et al. Fetal size and dating: charts recommended for clinical obstetric practice. *Ultrasound* 2009; 17: 161–167.
- Keraudren K, Kuklisova-Murgasova M, Kyriakopoulou V, et al. Automated fetal brain segmentation from 2D MRI slices for motion correction. *Neuroimage* 2014; 101: 633–643.
- Hadlock FP, Harrist RB, Sharman RS, et al. Estimation of fetal weight with the use of head, body, and femur measurements – a prospective study. *Am J Obstet Gynecol* 1985; 151: 333–337.
- Martins WP and Nastri CO. Interpreting reproducibility results for ultrasound measurements. *Ultrasound Obstet Gynecol* 2014; 43: 479–480.
- Kehl S, Schmidt U, Spaich S, et al. What are the limits of accuracy in fetal weight estimation with conventional biometry in two-dimensional ultrasound? A novel postpartum study. *Ultrasound Obstet Gynecol* 2012; 39: 543–548.
- Dudley N, Russell S, Ward B, et al.; BMUS QA Working Party. BMUS guidelines for the regular quality assurance testing of ultrasound scanners by sonographers. *Ultrasound* 2014; 22: 8–14.
- Sarris I, Ioannou C, Dighe M, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. *Ultrasound Obstet Gynecol* 2011; 38: 681–687.
- PHAST/DoH. Errors in epidemiological measurement. Health Knowledge – Public Health Action Support Team, 2011. Available at <https://www.healthknowledge.org.uk/e-learning/epidemiology/practitioners/errors-epidemiological-measurements>.
- Nesbitt Hawes EM, Tetstall E, Gee K, et al. Ultrasound (in)accuracy: it's in the formulae not in the technique – assessment of accuracy of abdominal circumference measurement in term pregnancies. *Australas J Ultrasound Med* 2014; 17: 38–44.
- Kyriakopoulou V, Vatansever D, Davidson A, et al. Normative biometry of the fetal brain using magnetic resonance imaging. *Brain Struct Function* 2017; 222: 2295–2307.
- Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. *Ultrasound Obstet Gynecol* 2005; 25: 80–89.
- Schild RL. Three-dimensional volumetry and fetal weight measurement. *Ultrasound Obstet Gynecol* 2007; 30: 799–803.
- Bennini JR, Marussi EF, Barini R, et al. Birth-weight prediction by two- and three-dimensional ultrasound imaging. *Ultrasound Obstet Gynecol* 2010; 35: 426–433.
- Lima JC, Miyague AH, Filho FM, et al. 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.
- National Institute of Health and Clinical Excellence *Antenatal care for uncomplicated pregnancies*. London, UK. NICE guideline (CG62). 2008.
- Dudley NJ. A systematic review of the ultrasound estimation of fetal weight. *Ultrasound Obstet Gynecol* 2005; 25: 80–89.
- Hatab MR, Zaretsky MV, Alexander JM, et al. Comparison of fetal biometric values with sonographic and 3D reconstruction MRI in term gestations. *Am J Roentgenol* 2008; 191: 340–345.
- Uotila J, Dastidar P, Heinonen T, et al. Magnetic resonance imaging compared to ultrasonography in fetal weight and volume estimation in diabetic and normal pregnancy. *Acta Obstet Gynecol Scand* 2000; 79: 255–259.
- Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; 64: 96–106.