Article

Quantitative Approach to Quality Review of Prenatal Ultrasound Examinations: Estimated Fetal Weight and Fetal Sex

C. Andrew Combs [*] , Ryan C Lee , Sarah Y Lee , Sushma Amara , Olaide Ashimi Balogun

Article

Quantitative Approach to Quality Review of Prenatal Ultrasound Examinations: Estimated Fetal Weight and Fetal Sex

C. Andrew Combs [1],*, Ryan C. Lee [2], Sarah Y. Lee [2], Sushma Amara [3] and Olaide Ashimi Balogun [4]

[1] Pediatrix Center for Research, Education, Quality & Safety, Sunrise, Florida, USA

[2] Obstetrix of California, Campbell, California, USA; ryan.c.lee@gmail.com

[3] Eastside Maternal-Fetal Medicine, Bellevue, Washington, USA

[4] Obstetrix Maternal-Fetal Medicine Specialists, Houston, Texas, USA

* Correspondence: andrew.combs@pediatrix.com Tel: +1 (408) 314-1792

Abstract: Background/Objectives: Systematic quality review of ultrasound exams is recommended to assure accurate diagnosis. Our primary objectives were to develop a quantitative method for quality review of estimated fetal weight (EFW) and to assess accuracy of EFW for an entire practice and for individual personnel. A secondary objective was to evaluate accuracy of fetal sex determination. **Methods:** This is a retrospective cohort study. Eligible ultrasound exams included singleton pregnancies with live birth and known birth weight (BW). A published method was used to predict BW from EFW for exams with ultrasound-to-delivery intervals up to 12 weeks. Mean error and median absolute error (AE) were compared between different personnel. Image audits were performed for exams with AE>30% and exams with reported fetal sex different than newborn sex. **Results:** We analyzed 1,938 exams from 890 patients. In the last exam before birth, median AE was 5.9% and predicted BW was within ±20% of actual BW in 97.2% of patients. AE was >30% in 28 exams (1.4%); image audit found correct caliper placement in all 28. Only 2 patients (0.2%) had AE>30% on the last exam before birth. One sonographer systematically over-measured head and abdominal circumferences, leading to EFWs that were overestimated. Reported fetal sex differed from newborn sex in 7 exams (0.4%), 5 patients (0.6%). Images in 4 of these patients were annotated with correct fetal sex but a clerical error was made in the report. In 1 patient, the unclear image was labeled "probably female" but the newborn was male. **Conclusions:** Accuracy of EFW in this practice was similar to literature reports. Quantitative analysis identified a sonographer with outlier measurements. Time-consuming image audits could be focused on a small number of exams with large errors. We suggest some enhancements to ultrasound reporting software that may help to reduce clerical errors. We provide tools to help other practices perform similar quality reviews.

Keywords: diagnostic error; fetal biometry; fetal growth restriction; large for gestational age; small for gestational age

1. Introduction

Estimation of fetal weight is an essential component of the standard obstetrical examination in the 2nd and 3rd trimesters of pregnancy [1]. Estimated fetal weight (EFW) derived from ultrasound biometry is central to the diagnosis of abnormal fetal growth, both fetal growth restriction (FGR) and large-for-dates. However, sonographic EFW is known to be imperfect; over 20% of EFWs differ from BW by more than 10% [2]. Such a high rate of relatively large errors results in a high potential for misdiagnosis of abnormal fetal growth.

Misdiagnosis of abnormal fetal growth has clinical consequences. For pregnancies with FGR, fetal surveillance is recommended, including serial assessments of fetal well-being, umbilical artery Doppler ultrasound, and fetal growth; with severe FGR, early delivery is recommended [3]. These interventions all have both direct financial costs and the indirect costs of patient time, inconvenience, and loss of work. For pregnancies suspected to be large-for-dates, induction of labor is sometimes

recommended and the rate of cesarean delivery is increased, even if actual birth weight (BW) is not increased [4–12].

Various professional organizations recommend systematic quality review to assure the accuracy of obstetrical ultrasound diagnoses [13–16]. We recently presented a method for a quantitative review of fetal biometric measurements, comparing the z-scores of measurements obtained by individual sonographers within a practice [17]. We demonstrated how image audits could be focused on sonographers whose mean z-scores deviated from the practice-wide mean, thereby saving the time and expense of auditing those whose measurements fell within the expected range. If sonographers vary in the accuracy of their biometric measurements, it seems likely that they might also vary in the accuracy of their EFWs.

The primary aim of the present study was to develop and demonstrate methods to assess the accuracy of EFW for an ultrasound practice overall and for individual providers. A secondary aim was to assess the accuracy of fetal sex determination in prenatal ultrasound exams.

2. Materials and Methods

2.1. Study design, setting

A retrospective quality review was undertaken for the cohort of ultrasound exams performed from January 2022 through December 2023 at Obstetrix of California, a suburban, referral-based maternal-fetal medicine (MFM) practice. During the study period, the practice used GE Voluson ultrasound machines (GE Healthcare, Wauwatosa, Wisconson, USA) with variable-frequency curvilinear transducers (2 to 9 mHz as needed to optimize imaging). We used GE Viewpoint software (Version 6) for generating reports and for storing images and metadata. Before starting the study, the protocol was reviewed by the Institutional Review Board of Good Samaritan Hospital (GSH, San Jose, California USA). Retrospective study of existing data was determined to be of minimal risk to patients and was granted Exempt status (#FWA-00006889, October 11, 2023).

2.2. Data extraction, inclusion and exclusion criteria

Using the Viewpoint query tool, we extracted data for all ultrasound exams meeting these eligibility criteria: fetal cardiac activity present and 4 basic fetal measurements recorded: biparietal diameter (BPD), head circumference (HC), abdominal circumference (AC), and femur length (FL). The extracted data for each exam included patient identifiers (name, date of birth, person number), exam identifiers (date, exam number), sonographer who performed the exam, physician who interpreted it, referring provider name, exam date, best obstetrical estimate of due date (EDD) [17], fetal measurements, and fetal sex.

We divided patients into 3 groups based on the referring obstetrical provider: (1) those with likely delivery at GSH because the referring physician had delivery privileges there; (2) those with possible delivery at GSH because the provider was from a distant hospital without neonatal intensive care, possibly reflecting a maternal transport; and (3) those unlikely to deliver at GSH because the provider practiced primarily at another hospital. From the first 2 groups, we created a spreadsheet containing patient identifiers and EDD but no other exam data. An investigator reviewed the GSH electronic record system to locate the patient and to record the delivery date, birth weight (BW), newborn sex, mode of delivery, admission to neonatal intensive care (yes/no), any congenital anomalies identified at birth, and the self-reported maternal race. This review was performed without access to any of the ultrasound findings. The file with delivery data was then merged with the data file containing the ultrasound exam information.

For analysis, we excluded multifetal pregnancies, stillbirths, and cases with gastroschisis or omphalocele. We excluded cases with extreme outliers for HC, AC, or FL (values more than 6 standard deviations (SD) from the normal mean based on formulas derived from World Health Organization norms [18–21]. We also excluded deliveries with an implausible combination of EDD and delivery date, i.e., a calculated gestational age at birth <0 or >50 weeks; most of these were attributed to a repeat pregnancy for which we did not have ultrasound data.

For patients with more than 1 eligible exam during the study period, all exams were included. Separate analyses were restricted to only the last exam, i.e., the shortest interval from ultrasound to delivery (latency.

2.3. Calculation of EFW, errors, and predicted BW

EFW was calculated from HC, AC, and FL using the 3-parameter formula of Hadlock et al [22] because 2 large comparative reviews concluded that this is the most accurate of over 70 formulas for calculating EFW [2,23].

Initial analysis was focused on exams performed within 1 day before birth, where accuracy of EFW was assessed by comparing it to birth weight (BW). Taking EFW as the predicted BW ($BW_{predicted}$), with error expressed as a percentage:

Error = $100 \cdot (BW_{predicted} - BW)/BW$

Absolute Error = $|Error|$

where $\cdot$ signifies multiplication. Positive values of error reflect overestimation of EFW and negative values reflect underestimation. Absolute error reflects only the magnitude of the error, ignoring its direction.

Because the number of ultrasound exams within 1 day of birth was too small for meaningful comparisons between sonographers, we used the method of Mongelli and Gardosi [24] to predict BW from EFW in exams with latencies of several weeks. The method predicts BW based on the assumption that the ratio of EFW to median normal fetal weight at the gestational age of the ultrasound exam (GA_{us}) stays constant over time and should therefore equal the ratio of BW to the median normal fetal weight at the gestational age of birth (GA_{birth}). This can be simplified to the formula:

$BW_{predicted} = EFW \cdot e^x$

where $x = -0.00354 \cdot (GA_{birth}^2 - GA_{us}^2) + 0.332 \cdot (GA_{birth} - GA_{us})$

where GA_{us} and GA_{birth} are expressed in exact weeks and e is the Euler constant (approximately 2.718). The derivation and validation of this formula are given in Supplementary File 1. We also show in Supplementary file 1 that the assumption of constant weight ratio is equivalent to assuming that the EFW percentile and z-core remain constant from ultrasound to birth. As shown in Supplementary File 1, z-score of EFW is highly correlated with z-score of BW for latencies <12 weeks (r=0.82, 0.71, 0.66 at latency of 0-3.9 weeks, 4-7.9 weeks, and 8-11.9 weeks, respectively). Because of decreasing accuracy with increasing latency, exams with latencies ≥12 weeks were excluded from the analysis.

2.4. Statistical analyses

Analyses were performed with Stata (version 18, Statacorp, College Station, TX, USA). Two-tailed P-values less than 0.05 were considered significant.

Errors were normally distributed and are presented as mean ± standard deviation (SD). Between-group errors were compared using one-way analysis of variance (ANOVA) with Sidak multiple-comparisons test. Absolute errors were not normally distributed and are presented as median with interquartile range (IQR). Between-group comparisons of absolute error were made with Kruskal-Wallis and U-tests. We compared errors and absolute errors between exams of differing latency (in 4-week blocks) and between individual sonographers and physicians.

Several exploratory analyses were performed to evaluate possible factors that might influence the accuracy of $BW_{predicted}$ including maternal race, maternal obesity, maternal age, newborn sex, gestational age at ultrasound or birth, birth weight, or fetal weight percentile. These are presented in Supplementary File 2.

2.5. Audit of exams with large errors

The protocol specified that audits would be performed for exams with absolute error more than 40%. We expanded this to audit all exams with absolute error more than 30% because the number of such exams was small. Each audit started with a manual double-check of the hospital record to confirm that the BW and delivery date were entered correctly in our data file. Then we reviewed

ultrasound reports and images to assess whether there were measurement inaccuracies, errors in transcription, or other factors to explain the error.

Similar audits were performed for all cases in which fetal sex on the ultrasound exam report was different than the newborn sex recorded in the delivery records and for the one case excluded because of an extreme outlier biometric value.

3. Results

3.1. Summary of included and excluded exams

A total of 1,938 ultrasound examinations in 890 patients met all the inclusion criteria and were included in the analyses. The flow diagram in Figure 1 shows the progress of exams from initial eligibility to final inclusion.
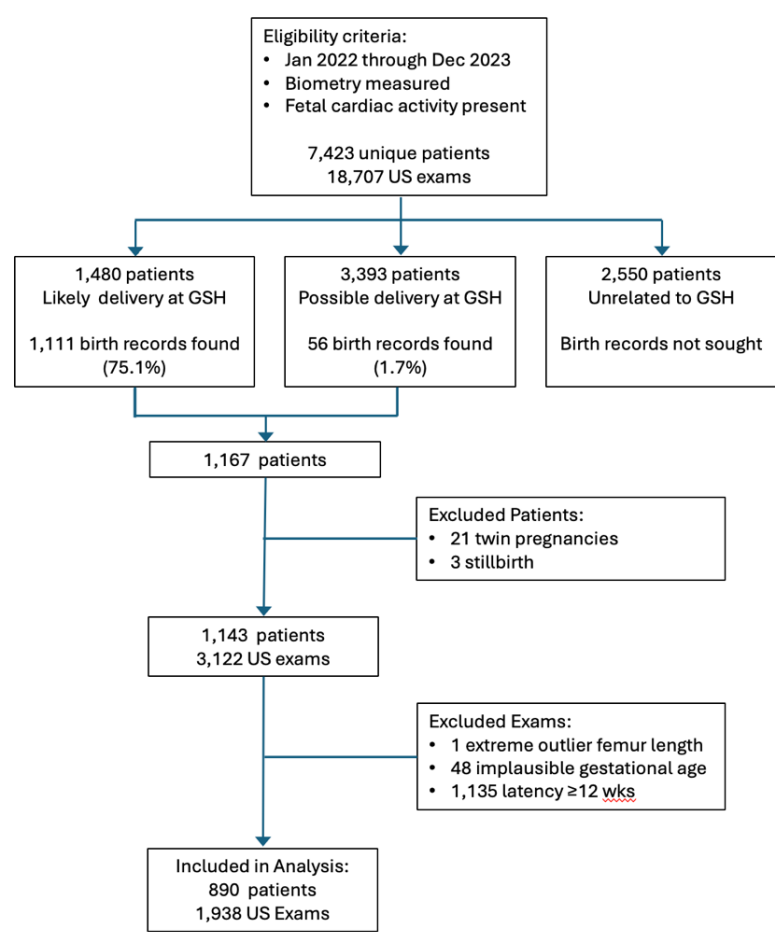

Figure 1. Flow diagram showing numbers of patients, eligible exams, exclusions, and exams included in final analysis. Abbreviations: GSH – Good Samaritan Hospital. US – ultrasound.

3.2. Accuracy of EFW for exams within 1 day of birth

Ultrasound was performed within one day of birth in 53 patients. For these, birthweight averaged $2,819 \pm 782$ gm (mean $\pm$ SD) at gestational age 36.2 ± 3.4 weeks. The difference between EFW and BW averaged -88 ± 266 gms (mean $\pm$ SD) for a mean error of $-2.3 \pm 8.8\%$ of BW. The negative value of error signifies that EFW tended to be smaller than BW but the mean error was not significantly different than 0 (P=0.07, t-test). Median absolute error was 6.3% (IQR 3.1- 9.0%). Absolute errors less than 10% were seen in 41 cases (77%), less than 20% in 50 cases (94%), and 20-30% in 3 cases (6%). These 53 exams were performed by 17 sonographers who performed from 1 to 5 exams each, so meaningful comparisons between sonographers was not feasible. No sonographer had more than 1 exam with absolute error >20%.

3.2. Accuracy of predicted birthweight for exams within 12 weeks of birth

The accuracy of $BW_{predicted}$ at different latencies is summarized in Table 1. In every 4-week latency block, $BW_{predicted}$ was higher than BW by about 3-5%; this overestimation was significant in every block and for the overall sample (P<0.001, t-test). In every latency block, median absolute error was ≤7%. Overall, over 71% of predictions were within 10% of BW and over 22% of predictions were within 10-20% of BW, so 93.9% of predictions were within 20% of BW. Even with latency of 8-11.9 wks, 89.8% of predictions were within 20% of BW.

Table 1. Accuracy of birth weight predictions at different latencies for entire practice.

Latency	N	Percent Error, mean ± SD	Percent Absolute Error, Median (IQR)	Exams with Absolute Error less than 10%, n (%)	Exams with Absolute Error 10 to <20%, n (%)	Exams with Absolute Error 20 to <30%, n (%)	Exams with Absolute Error 30% or more, n (%)
All Exams							
0-3.9 wks	800	2.9 ± 8.7[a]	5.9 (3.1-9.8)	609 (76.1%)	165 (20.6%)	21 (2.6%)	5 (0.6%)
4-7.9 wks	706	3.8 ± 10.0[a]	6.4 (3.0-11.3)	490 (69.4%)	169 (23.9%)	37 (5.2%)	10 (1.4%)
8-11.9 wks	432	4.8 ± 11.4[ab]	7.0 (3.3-11.9)[c]	288 (66.7%)	100 (23.2%)	31 (7.2%)	13 (3.0%)
Total	1,938	3.7 ± 9.9[a]	6.4 (3.1-11.0)	1386 (71.5%)	435 (22.5%)	89 (4.6%)	28 (1.4%)
Last Exam Before Birth							
0-3.9 wks	691	2.8 ± 8.5[a]	5.9 (3.1-9.8)	525 (76.0%)	148 (21.4%)	16 (2.3%)	2 (0.3%)
4-7.9 wks	155	2.5 ± 8.7[a]	5.7 (2.7-10.2)	116 (74.8%)	35 (22.6%)	4 (2.65)	0
8-11.9 wks	44	2.7 ± 9.8	5.9 (2.8-9.4)	33 (75.0%)	8 (18.2%)	3 (6.8%)	0
Total	890	2.8 ± 8.6[a]	5.9 (3.0-9.9)	674 (75.7%)	191 (21.5%)	23 (2.5%)	2 (0.2%)

Latency is the interval from ultrasound exam to birth, [a] significantly different than 0, t-test, P<0.001, [b] significantly different than 0-3.9 week group, ANOVA with Sidak test, P<0.005, [c] significantly different than 0-3.9 week group, u-test, P<0.005.

As shown in the lower half of Table 1, when analysis was restricted to the last ultrasound exam before birth, there was lower mean error, lower median percent error, and a 97.2% rate of predictions were within 20% of BW. Moreover, there was no significant difference in accuracy between different latency blocks in this subset, that is, exams with latency 8-11.9 weeks had similar accuracy to those with latency 0-3.8 weeks.

3.3. Accuracy of predicted birth weight for individual sonographers

The accuracy of predicted BW for 7 individual sonographers is summarized in Table 2. To minimize the possibility of spurious values with small sample size, we have restricted the Table to sonographers with more than 100 EFWs during the measurement period. All these sonographers had mean errors significantly above 0 ($BW_{predicted}$ higher than BW). Sonographer 9 had significantly larger mean error and mean absolute error than other sonographers.

Table 2. Accuracy of birth weight predictions for sonographers with at least 100 exams.

Sonographer Number	N	Percent Error, mean ± SD	Percent Absolute Error, median (IQR)	Exams with Absolute Error less than 10%,	Exams with Absolute Error 10 to <20%, n (%)	Exams with Absolute Error 20 to <30%, n (%)	Exams with Absolute Error 30% or more, n (%)

				n (%)			
All Exams							
6	212	5.3 ± 8.6 [a]	6.8 (3.3-11.4)	147 (70.0%)	51 (24.3%)	11 (5.2%)	1 (0.5%)
8	225	2.5 ± 9.5 [a]	6.3 (2.9-11.2)	157 (70.1%)	58 (25.9%)	7 (3.1%)	2 (0.9%)
9	127	6.8 ± 11.2 [ab]	8.3 (3.6-14.5) [c]	81 (62.8%)	34 (26.4%)	9 (7.0%)	5 (3.9%)
16	154	3.1 ± 9.4 [a]	6.2 (3.3-9.5)	118 (77.6%)	27 (17.8%)	5 (3.3%)	2 (1.3%)
17	145	2.7 ± 9.4 [a]	5.3 (2.8-9.3)	115 (79.3%)	23 (15.9%)	2 (1.4%)	5 (3.5%)
22	267	3.5 ± 10.8 [a]	6.3 (2.6-11.2)	190 (71.2%)	61 (22.9%)	11 (4.1%)	5 (1.9%)
24	225	5.4 ± 9.5 [a]	6.7 (3.2-11.3)	158 (70.2%)	49 (21.8%)	17 (7.6%)	1 (0.4%)
Total	1,938	3.7 ± 9.9 [a]	6.4 (3.1-11.0)	1386 (71.5%)	435 (22.5%)	89 (4.6%)	28 (1.4%)
Last Exam Before Birth							
6	90	5.0 ± 7.6 [a]	6.4 (3.3-9.8)	68 (76%)	20 (22%)	2 (2%)	0
8	105	2.0 ± 8.6 [a]	6.0 (3.3-10.2)	77 (73%)	27 (26%)	0	1 (1%)
9	44	6.0 ± 8.9 [ab]	7.6 (3.5-11.8) [c]	31 (70%)	11 (25%)	2 (5%)	0
16	76	2.1 ± 7.3 [a]	5.8 (3.3-8.4)	61 (80%)	15 (20%)	0	0
17	58	1.4 ± 7.9	5.0 (3.3-9.1)	47 (91%)	10 (17%)	1 (2%)	0
22	105	2.3 ± 8.3 [a]	5.5 (2.6-9.8)	81 (77%)	22 (21%)	2 (2%)	0
24	84	5.2 ± 7.9 [a]	5.9 (3.1-9.4)	67 (80%)	12 (14%)	5 (6%)	0
Total	890	2.8 ± 8.6 [a]	5.9 (3.0-9.9)	674 (75.7%)	191 (21.5%)	23 (2.5%)	2 (0.2%)

Total includes 17 additional sonographers with ≤100 exams each. a – significantly different than 0, t-test, P<0.05, b – significantly different than Sonographer 8, Sidak test, P=0.005, c – significantly different than Sonographers 16 and 17, Kruskal Wallace test, P<0.01.

The lower half of Table 2 summarizes the accuracy of the last exam before birth. Mean error and median absolute error were smaller in this subset than in the All Exams cohort and 97.2% of predictions were within 20% of BW. Although sonographer 9 still had higher mean error and absolute error than the others, the difference did not reach significance in this subset.

We considered two possible reasons why sonographer 9 had larger mean error than the others: first, if the exams had longer latency interval between ultrasound and birth than other sonographers' exams, a larger error might be expected; and second, if fetal biometry was systematically over-measured, the EFWs would have larger error. The mean latency and biometry z-scores for the 7 sonographers with at least 100 exams each are summarized in Table 3. Mean latency for sonographer

9 was slightly higher than for the others, but not significantly so. Using the z-score method for assessment of biometry that we previously described [25], Sonographers 9 and 24 systematically overmeasured HC and AC (z-scores significantly larger than the practice mean). Sonographers 16 and 22 also systematically overmeasured HC. Sonographers 9 and 24 are no longer with the practice, so no further audit was done for them.

Table 3. Latency and biometry for sonographers with at least 100 exams.

Sonographer Number	N	Latency, weeks mean ± SD	Head Circumference z-score, mean ± SD	Abdominal Circumference z-score, mean ± SD	Femur Length z-score, mean ± SD
6	212	5.1 ± 3.0	0.09 ± 1.00	0.19 ± 0.82	0.33 ± 1.07
8	225	5.6 ± 3.1	−0.10 ± 0.88	0.22 ± .80	0.16 ± 0.99
9	127	6.0 ± 3.3 [a]	0.18 ± 0.85 [b]	0.38 ± 0.74 [b]	0.25 ± 1.00
16	154	4.9 ± 3.2	0.21 ± 1.04 [b]	0.20 ± 0.79	0.09 ± 1.00
17	145	5.6 ± 3.1	0.02 ± 0.92	0.18 ± 0.72	0.09 ± 0.67
22	267	5.6 ± 3.2	−0.22 ± 1.03 [b]	0.15 ± 0.78	0.45 ± 1.10 [bd]
24	225	5.7 ± 3.2	0.13 ± 0.92 [b]	0.47 ± 0.85 [bc]	0.27 ± 0.86
Practice Total	1,938	5.2 ± 3.2	−0.01 ± 0.98	0.21 ± 0.82	0.17 ± 1.01

Practice Total includes 17 additional sonographers with ≤100 exams each. [a] P<0.05 for overall between-sonographer difference (ANOVA), P=0.08 comparing sonographers 9 vs 16 (Sidak test). [b] P<0.05 compared to practice mean (t-test) [c] P <0.05 compared to Sonographers 6, 8, 15, 17, and 22 (ANOVA with Sidak test) [d] P <0.05 compared to Sonographers 8, 16, and 17 (ANOVA with Sidak test).

3.4. Accuracy of predicted BW for individual physicians

In most cases, biometry measurements are obtained by sonographers. If multiple measurements of the same biometric parameter are obtained, the ultrasound machine will automatically report the average measurement to the Viewpoint software. Physicians may sometimes repeat the measurements or may select a particular measurement for reporting, overriding the automated average. If a physician does this repeatedly, it is possible that their EFWs will be more accurate or less accurate.

Table 4 evaluates the possibility that physicians might differ in the accuracy of their BW predictions. Physician 6 had slightly lower mean error than the others but no significant difference in median absolute error. The difference in mean error was still evident but no longer significant in last exam before birth. Physician 6 is known to change the sonographers' measurements more often than the other physicians and this appears to have been associated with a slight improvement in accuracy:

3.5. Exploration of other factors potentially affecting accuracy of predicted BW

Supplementary File 2 contains analyses of the accuracy of BW predictions stratified by maternal race, maternal obesity, maternal age, newborn sex, gestational age, and fetal weight. Some of these factors had statistically significant associations with accuracy (e.g., mean error was slightly lower in Hispanic patients than in Asian or White patients, lower in male than female fetuses, and higher with ultrasound exams before 30 weeks of gestation; absolute error was higher among preterm births) but none of these was large enough to have major bearing on the quality review. We found no association between accuracy and maternal obesity or maternal age.

Table 4. Accuracy of birth weight predictions by interpreting physicians.

Physician Number	N	Percent Error,	Percent Absolute Error,	Exams with Absolute Error	Exams with Absolute	Exams with Absolute	Exams with Absolute

		mean ± SD	Median (IQR)	less than 10%, n (%)	Error 10 to <20%, n (%)	Error 20 to <30%, n (%)	Error 30% or more, n (%)
All Exams							
1	164	4.8 ± 10.6 [a]	7.0 (3.0-11.5)	113 (68.9)	38 (23.2)	9 (5.5)	4 (2.4)
2	88	5.2 ± 9.1 [a]	6.8 (2.9-12.6)	61 (69.3)	24 (27.3)	2 (2.3)	1 (1.1)
3	351	3.3 ±10.2 [a]	6.4 (3.6-10.9)	245 (69.8)	85 (24.2)	14 (4.0)	7 (2.0)
4	448	3.8 ± 9.7 [a]	6.3 (3.0-11.3)	324 (72.3)	94 (31.0)	27 (6.0)	3 (0.7)
5	415	4.3± 9.7 [a]	6.5 (3.0-11.3)	297 (71.6)	93 (22.4)	20 (4.8)	5 (1.2)
6	472	2.5 ± 9.7 [ab]	6.0 (2.9-10.7)	346 (73.3)	101 (21.4)	17 (3.6)	8 (1.7)
Total	1,938	3.7 ± 9.9[a]	6.4 (3.1-11.0)	1386 (71.5%)	435 (22.5%)	89 (4.6%)	28 (1.4%)
Last Exam Before Birth							
1	72	2.6 ± 9.1 [a]	5.4 (2.5-10.6)	53 (73.6)	18 (25.0)	0	1 (1.4)
2	38	4.5 ± 6.8 [a]	5.1 (2.9-9.3)	29 (76.3)	9 (23.7)	0	0
3	170	2.5 ± 8.4 [a]	6.3 (4.1-10.4)	123 (72.4)	44 (25.9)	3 (1.8)	0
4	193	3.0 ± 8.9 [a]	5.9 (2.9-9.6)	153 (79.3)	31 (16.1)	9 (4.7)	0
5	203	3.5 ± 8.6 [a]	6.1 (2.8-10.0)	153 (75.4)	42 (20.7)	8 (3.9)	0
6	214	1.8 ± 8.4 [a]	5.7 (2.5-9.5)	163 (76.2)	47 (22.0)	3 (1.4)	1 (0.5)
Total	890	2.8 ± 8.6 [a]	5.9 (3.0-9.9)	674 (75.7%)	191 (21.5%)	23 (2.5%)	2 (0.2%)

Total includes 17 additional sonographers with ≤100 exams each. a – significantly different than 0, t-test, P<0.05b – significantly different than Physician 5, Sidak test, P=0.05c – significantly different than Sonographers 16 and 17, Kruskal Wallace test, P<0.01.

3.6. Diagnostic accuracy of prediction of small- or large-for gestational age (SGA or LGA)

Supplementary File 3 tabulates the test performance characteristics of using FGR (defined as either EFW or AC <10th percentile to predict SGA (BW <10th percentile) or using EFW >90th percentile to predict LGA (BW >90th percentile). These analyses were based on the last exam before birth.

Diagnosis of FGR had low sensitivity (51%) but high specificity (92%) for the prediction of SGA. When analysis was restricted to exams <7 days before birth, sensitivity was improved (84%) but specificity was lower (83%). The areas under the receiver operating characteristic (AUROC) curves relating EFW percentile to SGA were 0.88 and 0.93, respectively.

EFW >90th percentile had low sensitivity and high specificity for the prediction of LGA, irrespective of whether analysis included all exams (43%, 97%, respectively) or only exams <7 days before birth (44%, 95%). The AUROCs were 0.91 and 0.95 respectively.

3.7. Audit of exams with large errors

We audited the single exam that was excluded because of an extreme outlier biometric value (FL more 14 SD below the mean). On this exam, the FL was recorded as 3.2 mm but the images showed FL 3.21 cm (i.e., a 10-fold error).

We audited all exams with absolute error more than 30% (28 exams in 17 patients). Image audit found acceptable caliper placement for biometric measurements in all 28 of these exams. In 15 patients, subsequent exams showed smaller absolute errors, so absolute errors >30% were seen on the last exam before birth in only 2 of 890 patients (0.2%). For 14 of the other 15 patients, $BW_{predicted}$ was higher than BW (errors >30%) on an early exam but subsequent exams showed progressively smaller errors; fetal growth restriction was ultimately diagnosed in 10 of these patients. For the 15th patient, $BW_{predicted}$ was lower than BW (error < –30%); this was an ultrasound exam at 26 weeks of gestation in a patient with Type 1 diabetes; $BW_{predicted}$ was 3021 gm but BW was 4620 gm at 37 weeks (latency 11 weeks).

3.7. Accuracy of fetal sex reporting

Differences between newborn sex and fetal sex recorded with the ultrasound exam were found in 7 of 1,938 exams (0.4%), 5 of 890 patients (0.6%), as summarized in Table 5. Audit found that the correct fetal sex was seen and labeled on the images for Cases 1 through 4. For one patient (Case 2), the error reported on the first exam was carried forward to two later exams even though the genital area was not imaged on the follow-up exams. In one exam (Case 5), the image was unclear but labeled as "probably female" and recorded as female in the database; fetal sex was not disclosed to patient or included in the report to the referring provider, at patient request; the newborn was male.

Table 5. Cases with recorded fetal sex different than newborn sex.

Case Number	Sonographer Number	Gestational Age at Ultrasound, weeks + days	Fetal Sex Reported	Newborn Sex	Review Findings
1	16	36+5	Female	Male	Prior exam reported male. No images of genital area on repeat exam; no explanation why sex was changed or why reported.
2.1	8	28+5	Male	Female	Images correctly labeled female, error in database and on report.
2.2	8	32+2	Male	Female	No images of genital area, sex carried over from prior exam.
2.3	17	36+3	Male	Female	No images of genital area, sex carried over from prior exam.
3	13	35+1	Female	Male	Images correctly labeled male, error on report.
4	6	32+4	Male	Female	Images correctly labeled female, error on report.
5	7	30+6	Female	Male	Images labeled "probably female", recorded in database as female. Patient did not want to know sex, so no sex appeared on report.

In addition, fetal sex was recorded as "uncertain" in 13 exams (0.7%) from 11 patients (1.2%). These were performed by 7 different sonographers, with each having only 1-3 exams with uncertain fetal sex, a number too small to suggest a systematic problem. For 4 of these exams (3 patients), the correct sex was recorded in another exam of the same patient. In 4 exams (3 patients), the fetus was in breech presentation with GA_{us} >30 weeks, so fetal genitalia could not be seen. For the remaining 5 exams (5 patients), no specific reason was given for the uncertainty; 3 of these had other exams (not

included in the analysis because of long latency) where sex was also recorded as uncertain. None of the newborns had ambiguous genitalia.

4. Discussion

4.1. Principal findings

In this practice, EFW obtained up to 12 weeks before birth yielded accurate predictions of BW. In the last exam before birth, median absolute error was less than 6%, over 75% of predictions were within 10% of BW, and over 97% were within 20% of BW. These results are comparable to those reported in a study of 5,163 exams performed within 2 days of birth, where absolute error averaged 6.7% and 77.7% of EFWs were within 10% of BW [2].

Between-sonographer comparisons revealed one sonographer whose mean EFW was significantly overestimated compared to the others, attributable to overmeasurement of HC and AC. Between-physician comparisons revealed one physician whose mean EFW was slightly less overestimated than the others, but there was no significant difference in median absolute error. We are not aware of any prior studies comparing the accuracy of individual sonographers or physicians.

Audit of exams with large errors did not reveal any images with incorrect caliper placement. Only 2 patients of 890 (0.2%) had ≥30% error on the last exam before birth. In 1 excluded case with an extreme outlier measurement, the FL was recorded in centimeters rather than millimeters, a clerical error.

Fetal sex was incorrectly recorded in 0.4% of exams (0.6% of patients). In one case, the sonographer was uncertain about fetal sex and recorded an incorrect guess. In all other cases, the sonographer had correctly determined the fetal sex but a clerical error was made in documenting the findings.

4.2. Prediction of birthweight after long latency

The assumption that fetal weight percentile remains constant as the fetus grows yielded very accurate predictions of birth weight in most cases. Most of the exceptions (i.e., those with absolute errors ≥30%) were cases with progressive fetal growth restriction (n=14) or evolving fetal macrosomia with diabetes (n=1). When analysis was restricted to the last exam before birth, only 2 patients (0.2%) had an error ≥30%.

Despite low mean errors and median absolute errors, however, the diagnostic accuracy of abnormal fetal size (FGR or EFW >90th percentile) had low sensitivity for predicting abnormal BW (SGA or LGA).

Diagnostic performance for predicting SGA was similar in our results versus those recently reported in a series of 10,045 patients (sensitivity 51% vs 36%, specificity 92% vs 96%, respectively) [26]. Sensitivity was substantially higher in exams <7 days before birth (84%), though with decreased specificity (83%).

Diagnostic performance for predicting LGA was similar in our results versus those recently reported in a series of 26,627 exams (sensitivity 43% vs 41%, specificity 97% vs 95%, respectively) [27].

Ultimately, the reason for diagnosing abnormal growth is not to predict whether BW will be abnormal, but whether the fetus will be at risk for morbidity [28]. To improve sensitivity for the prediction of risk of perinatal morbidity, different cut-off values than the traditional 10th and 90th percentiles may be needed, but different cut-offs may also result in decreased specificity [28].

We have shown that the prediction of BW after long latency is useful for quality review in that it allows for sample sizes sufficient to detect subtle inaccuracies in EFW determination. However, we do not encourage clinical diagnoses such as FGR or EFW >90th percentile to be made based on long-ago exams. Accuracy and diagnostic performance degrade with increasing latency, so clinical management decisions should be based only on an up-to-date EFW.

4.3. Quantitative analysis to guide focused image audit

The traditional approach to quality review of ultrasound exams involves auditing images for errors in measurement and interpretation. For example, in the RADPEER program of the American

College of Radiology (ACR), a percentage of examinations are randomly selected for "double-reading" by a second physician [29,30]. This approach is time-consuming and rarely uncovers clinically significant errors [30–33]. The RADPEER program is targeted primarily on structural findings rather than measurements. For fetal biometry, a similar method for quality monitoring was used by the INTERGROWTH-21st project in its international prospective study of fetal growth [34]. In this program, a 10% random sample of images was sent to a central Quality Unit for a second reading. This comprehensive method is laudable, but is likely too time-consuming and labor-intensive for a typical clinical practice.

In contrast, using our quantitative approach, the accuracy of EFW can be assessed without any reference to images at all. One needs only compare the actual BW to the BW predicted from EFW. While EFW, in turn, is derived from biometric measurements on the images, there is little value in auditing biometry images from the majority of exams in which EFW is reasonably accurate. Image audits can be targeted at the small number of exams with large errors.

We demonstrate how sonographers or physicians whose accuracy deviates from other personnel in the practice can be identified using the quantitative approach and standard statistical methods (Tables 2 and 4, respectively). For the one sonographer whose EFWs were systematically overmeasured, we demonstrate how a quantitative assessment of BPD, HC, AC, and FL can reveal systematic measurement deviations (Table 3) using methods we have described previously [25]. If the deviations are large enough, image audits can be targeted to those specific views.

The focused audits revealed several clerical errors. In a single case, FL was recorded in centimeters rather than millimeters; this error must have been made by someone entering the FL manually because the ultrasound machines transmit biometry measurements to the Viewpoint software automatically. Similarly, manual entry of fetal sex into the database was the likely cause in most of the cases of discordance between reported fetal sex and birth sex. Human error is a known contributor to misdiagnosis [35]. In one patient, the incorrectly assigned sex was carried forward automatically in subsequent exams even though fetal genitalia were not imaged again.

4.4. Measurement quality review in context

Review of the accuracy of EFW is only one component of a comprehensive quality program for prenatal ultrasound. Other components include review fetal biometry which we have previously described [Combs 2024] and review fetal anatomy imaging. In a forthcoming article, we will describe our quantitative approach to quality review for the fetal anatomy survey.

There are also several structural components generally recommended for a high-quality ultrasound practice [13–16]. These include accreditation of the practice by an organization such as the American Institute of Ultrasound in Medicine [15] or ACR [16]. Accreditation requires that all personnel have formal initial education and training, continuing education, certification or licensing. Accreditation also requires written protocols to ensure uniformity of exams, timely communication of results, cleaning and disinfection of transducers, equipment maintenance, and patient safety and confidentiality. An onboarding program is recommended for new personnel, including orientation to practice protocols and assessment of competency [13,14].

4.5. Strengths and limitations

A strength of the quantitative summary is that it provides a large-scale overview of an entire practice and accuracy of individual personnel (Tables 1-5). Analyzing a large number of exams for each provider, the method is highly sensitive to small between-provider variations, readily identifying outlier personnel for focused review. Once the Stata script is written, the quantitative analysis can be performed rapidly and repeated periodically. For those who wish to use these methods, we have provided a sample Excel spreadsheet with pseudodata (Supplementary File 4) and sample Stata scripts that will perform the basic analyses for Tables 1-5 (Supplementary Files 5 and 6).

An important limitation for our practice was that we needed to retrieve BW and other delivery data by manual review of the maternal hospital record, a time-consuming endeavor. We are not part of an integrated health system with an electronic health record system that automatically links

maternal and newborn information. Only 20% of our potentially eligible patients were from referring providers who deliver routinely at our primary hospital and we were only able to find delivery information on 75% of those (Figure 1). When we were unable to match patients by full name and date of birth, we were sometimes successful matching first name plus date of birth because surnames frequently change during the course of a pregnancy but first names generally do not. Some patients probably delivered elsewhere, but we doubt that this would account for our failure to match 25% of patients. All told, it took about 50 person-hours to perform the hospital record search and review. Less time would certainly have been required if our record system was integrated with the hospital's or if an automated database retrieval program had been available.

Another limitation is that some sonographers had a relatively small number of exams, even though the review period was 2 full years. We reviewed their results but did not include them in the Tables because small N's can produce spurious results.

Conversely, a large number of exams can result in very low p-values even with relatively small errors. For example, Sonographer 8 had a mean error of 2.0% in the last exam before birth (Table 2, lower half), which was significantly different than 0 ($BW_{predicted} > BW$, p=0.017); in a typical 3,500 gm newborn, an error of 2% is 70 gm, too small to be of clinical importance.

Another limitation is the inherent time lag between the ultrasound exam and the quality review, including both the latency from ultrasound to birth and the time it takes to accumulate sufficient cases for review. In the interim, there can be personnel turnover, as evidenced by the sonographers with outlier measurements who had left the practice before their systematic overmeasurement was detected.

4.6. Future directions – software enhancements

Viewpoint ultrasound reporting software includes data fields to record birth weight and delivery date. Other software may have similar fields. We suggest that software developers provide tools to link these newborn fields to prenatal ultrasound exams. Tools should also be developed to automate quality review of the accuracy of birth weight predictions from EFW for individual personnel and for the whole practice, including the mean error, median absolute error, and tabulation of exams with various degrees of error, similar to our Tables 1,2, and 4.

Minor software enhancements might help to prevent some clerical errors. First, findings from a prior exam should not automatically carry forward to subsequent exams. We found one patient for whom the same erroneous fetal sex was carried forward to 2 subsequent exams despite lack of any imaging of the fetal genitalia. On the other hand, useful information that can be carried forward from one exam to the next might include a reminder to reevaluate views that were suboptimal on the earlier exam, such as uncertainty about fetal sex or other elements of the anatomy, perhaps displaying incomplete items in red. Second, although it is important to allow manual override of automatically transferred biometrics, the reporting software should flag extreme outliers as a crosscheck, perhaps by displaying values in red. We had one case where an erroneous measurement was off by a factor of 10 (presumably manually entered), resulting in an EFW <1st percentile; this was not noticed and the report indicated "fetal size appropriate for dates."

5. Conclusions

We conclude that EFWs performed in this practice with latencies up to 12 weeks have accuracy similar to those reported in the literature for exams with latencies of only a few days. We identified one sonographer who systematically over-measured HC and AC, and whose EFWs therefore yielded $BW_{predicted}$ that was, on average, higher than actual BW. In practice-wide audit of the few exams with large errors, caliper placement was uniformly acceptable. Ultrasound diagnosis of fetal sex was correct in 99.9% of patients, but several clerical errors resulted in reports that were correct in only 99.4.

We believe it is important for every ultrasound practice to perform a quality review such as this. We have demonstrated a method and provided tools to allow practices to perform a similar evaluation.

Supplementary Materials: The following supporting information can be downloaded at: Preprints.org. S1: Derivation and validation of formula for birth weight prediction. S2: Exploration of factors potentially affecting accuracy of birth weight prediction S3: Diagnostic accuracy.S4: Sample Excel spreadsheet with pseudoexams for 1,000 observationsS5: Sample Stata analysis script for generating Tables 1 to 5 from spreadsheet (Word .docx format) S6: Sample Stata analysis script for generating Tables 1 to 5 from spreadsheet (Stata .do format).

Author Contributions: Conceptualization, CAC, RL, SL; methodology, CAC; software, CAC; validation, RL, SYL, OAB, SA; formal analysis, CAC; investigation, all authors; data curation, CAC; writing—original draft preparation, CAC; writing—review and editing, all authors; visualization, supervision, and project administration, CAC; funding acquisition, none. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. The protocol was reviewed by the Institutional Review Board (IRB) of Good Samaritan Hospital, San Jose. Because the study involved retrospective review of existing data and posed negligible risk to persons, it was determined to be Exempt.

Informed Consent Statement: Patient consent was waived due to the Exempt status of this retrospective analysis of existing data.

Data Availability Statement: For investigators wishing to develop their own analysis script, we provide a Supplemental Excel .XLSX file containing pseudodata with 1,000 observations and a sample Stata script file that can be used to generate Tables 1 to 5.

Conflicts of Interest: The Pediatrix Center for Research, Education, Quality & Safety (CAC) received funds from Sonio, Inc. for providing ultrasound images for training and validation of artificial intelligence software to detect fetal anomalies and enhance exam completeness. No artificial intelligence was used in the present studies or in preparing the manuscript for publication. Sonio had no role in the present studies.

References

1. American Institute of Ultrasound in Medicine. AIUM-ACR-ACOG-SMFM-SRU practice parameter for the performance of standard diagnostic obstetric ultrasound examinations. J Ultrasound Med 2018; 37:E13-E24.
2. Hammami A, Mazer Zumaeta A, Syngelaki A, Akolekar R, Nicolaides KH. Ultrasonographic estimation of fetal weight: development of new model and assessment of performance of previous models. Ultrasound Obstet Gynecol 2018; 52:35-43.
3. Society for Maternal-Fetal Medicine (SMFM); Martins JG, Biggio JR, Abuhamad A. Society for Maternal-Fetal Medicine Consult Series #52: Diagnosis and management of fetal growth restriction. Am J Obstet Gynecol 2020; 223(4): B2-B19.
4. Parry S, Severs CP, Sehdev HM, Macones GA, White LM, Morgan MA. Ultrasonic prediction of fetal macrosomia. Association with cesarean delivery. J Reprod Med 2000; 45:17-22.
5. Blackwell SC, Refuerzo J, Chadha R, Carreno CA. Overestimation of fetal weight by ultrasound: does it influence the likelihood of cesarean delivery for labor arrest? Am J Obstet Gynecol 2009; 200:340.e1-e3.
6. Melamed N, Yogev Y, Meizner I, Mashiach R, Ben-Haroush A. Sonographic prediction of fetal macrosomia. The consequences of false diagnosis. J Ultrasound Med 2010; 29:225-230.
7. Little SE, Edlow AG, Thomas AM, Smith NA. Estimated fetal weight by ultrasound: a modifiable risk factor for cesarean delivery? Am J JObstet Gynecol 2012; 207:309.e1-e6.
8. Yee LM, Grobman WA. Relationship between third-trimester sonographic estimation of fetal weight and mode of delivery. J Ultrasound Med 2016; 35:701-706.
9. Froehlich RJ, Gandoval G, Bailit JL, et al. Association of recorded estimated fetal weight and cesarean delivery in attempted vaginal delivery at term. Obstet Gynecol 2016; 128:487-494.
10. Matthews KC, Williamson J, Supta S, et al. The effect of a sonographic estimated fetal weight on the risk of cesarean delivery in macrosomic and small for gestational-age infants. J Matern Fetal Neonatal Med 2017; 30:1172-1176.
11. Stubert J, Peschel A, Bolz M, Glass A, Gerber B. Accuracy of immediate antepartum ultrasound estimated fetal weight and its impact on mode of delivery and outcome – a cohort analysis. BMC Pregnancy Childbirth 2018; 18:118.
12. Dude AM, Davis B, Delaney K, Yee LM. Sonographic estimated fetal weight and cesarean delivery among nulliparous women with obesity. Am J Perinatol Rep 2019; 9:e127-e132.

13. Benacerraf BF, Minton KK, Benson CB, et al. Proceedings: Beyond Ultrasound First Forum on improving the quality of ultrasound imaging in obstetrics and gynecology. Am J Obstet Gynecol 2018; 218:19-28.

14. Society for Maternal-Fetal Medicine. Executive summary: Workshop on developing an optimal maternal-fetal medicine ultrasound practice, February 7-8, 2023, cosponsored by the Society for Maternal-Fetal Medicine, American College of Obstetricians and Gynecologists, American Institue of Ultrasound in Medicine, American Registry for Diagnostic Medical Sonography, Internation Society of Ultrasound in Obstetrics and Gynecology, Gottesfeld-Hohler Memorial Foundation, and Perinatal Quality Foundation. Am J Obstet Gynecol 2023; 229:B20-4.

15. American Institute of Ultrasound in Medicine. Standards and guidelines for the accreditation of ultrasound practices. Available online: https://www.aium.org/resources/official-statements/view/standards-and-guidelines-for-the-accreditation-of-ultrasound-practices (accessed 28 May 2024).

16. American College of Radiology. Physician QA requirements: CT, MRI, nuclear medicine/PET, ultrasound (revised 1-3-2024). Available online: https://accreditationsupport.acr.org/support/solutions/articles/11000068451-physician-qa-requirements-ct-mri-nuclear-medicine-pet-ultrasound-revised-9-7-2021- (accessed 28 May 2024).

17. American College of Obstetricians and Gynecologists, Committee on Obstetric Practice, American Institute of Ultrasound in Medicine, Society for Maternal-Fetal Medicine. Method for estimating due date. Committee Opinion number 611. Obstet Gynecol 2014; 124(4):863-866.

18. Kiserud T, Piaggio G, Carroli G, Widmer M, Carvalho J, Jensen LN, Giordano D, Cecatti JG, Aleem HA, Talegawkar SA, et al. The World Health Organization Fetal Growth Charts: a multination study of ultrasound biometric measurements and estimated fetal weight. PLoS Med 2017; 14:e1002220.

19. Combs CA, Castillo R, Kline C, Fuller K, Seet E, Webb G, del Rosario A. Choice of standards for sonographic fetal abdominal circumference percentile. American Journal of Obstetrics and Gynecology Maternal-Fetal Medicine 4: 100732, 2022.

20. Combs CA, del Rosario A, Ashimi Balogun O, Bowman Z, Amara S. Selection of standards for sonographic fetal head circumference by use of z-scores. American Journal of Perinatology 41 (suppl 1): e2625-e2635, 2024.

21. Combs CA, del Rosario A, Ashimi Balogun O, Bowman Z, Amara S. Selection of standards for sonographic fetal femur length by use of z-scores. American Journal of Perinatology 41 (suppl 1): e3147-e3156, 2024.

22. 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.

23. Milner J, Arezina J. The accuracy of ultrasound estimation of fetal weight in comparison to birthweight: a systematic review. Ultrasound 2018; 24:32-41.

24. Mongelli M, Gardosi J. Gestation-adjusted projection of estimated fetal weight. Acta Obstet Gynecol Scand 1996; 75:28-31.

25. Combs

26. Leon-Martinez D, Lunsdberg LS, Culhane J, Zhang J, Son M, Reddy UM. Fetal growth restriction and small for gestational age as predictors of neonatal morbidity: which growth nomogram to use? Am J Obstet Gynecol 2023; 229:678.e1-e16.

27. Ewington LJ, Hugh O, Butler E, Quenby S, Gardosi J. Accuracy of antenatal ultrasound in predicting large-for-gestational-age babies: population-based cohort study. Am J Obstet Gynecol 2024; online ahead of print. DOI 10.1016/j.ajog.2024.04.052

28. Liauw J, Mayer C, Albert A, Fernandez A, Hutcheon JA. Which chart and which cut-point: deciding on the INTERGOWTH, World Health Organization, or Hadlock fetal growth chart? BMC Pregnancy Childbirth 2022:22:25.

29. Chaudhry H, Del Gaizo AJ, Frigini LA, et al. Forty-one million RADPEER reviews later: what we have learned and are still learning. J Am Coll Radiol 2020; 17:779-785.

30. Dinh ML, Yazdani R, Godiyal N, Pfeifer CM. Overnight radiology resident discrepancies at a large pediatric hospital: categorization by year of training, program, imaging modality, and report type. Acta Radiol 2022; 63:122-6.

31. Maurer MH, Bronnimann M, Schroeder C, et al. Time requirement and feasibility of a systematic quality peer review of reporting in radiology. Fortschr Rontgenstr 2021; 193:160-167.

32. Geijer H, Geijer M. Added value of double reading in diagnostic radiology, a systematic review. Insights Imaging 2018; 9:287-301.

33. Moriarity AK, Hawkins CM, Geis JR, et al. Meaningful peer review in radiology: a review of current practices and future directions. J Am Coll Radiol 2016; 13:1519-24.

34.   Cavallaro A, Ash ST, Napolitano R, et al. Quality control of ultrasound for fetal biometry: results from the INTERGROWTH-21st project. Ultrasound Obstet Gynecol 2018; 52:332-339.

35.   Kohn KT, Corrigan JM, Donaldson MS. To Err is Human: Building a Safer Health System. Washington, DC: National Academy Press, 1999.