The effect of preliminary training on quantitative evaluation of sonographer performance in the fetal morphology ultrasound examination

Abstract

Introduction: The aim of this study is to provide a quantitative scoring system to assess sonographer performance by reviewing images from the fetal morphology examination.

Methods: Ten ultrasound images from patients at 18-22 weeks gestation were assessed and scored for quality according to predefined criteria. One hundred normal cases were randomly selected and 10 images from each case were analysed by four experienced reviewers. The preliminary training incorporated the first 25 cases and involved a training period for reviewers; the remaining 75 cases were allocated to post training. The scores acquired by each reviewer were statistically analysed using Pearson’s and intra-class correlations to determine the reproducibility of the results.

Results: The preliminary training results were calculated separately and compared to the post training study. The preliminary intra-class correlation coefficient was 0.12. In the post training study the intra-class correlation coefficient was doubled at 0.24. The greatest correlation was observed between reviewers 1 and 4 with a coefficient of 0.71. Reviewers 3 and 4 demonstrated the lowest correlation coefficient of 0.30.

Discussion: A significant increase in the intra-class correlation coefficient indicated that training reviewers achieves more reproducible results. Suggested improvements to the study include recording fetal position, maternal BMI and assessing individual reviewer variability. An instruction manual defining each criterion might also yield better results.

Conclusion: The quantitative method used in this study assessed ultrasound images by placing a numerical value on image quality. Analysis of the preliminary training period demonstrates improved reproducibility of the results. Further investigation into the criteria is necessary to refine the quantitative method.

Keywords: audit, foetal, ultrasound, quality, screening score.
weeks gestation were assessed and scored according to quality. The images were of three biometrical and seven anatomical standardised ultrasound planes taken by six trained sonographers at Nepean Hospital a tertiary teaching hospital of the University of Sydney. The six sonographers were unaware of the study during the time of the ultrasound. A minimum of 1560 routine second trimester ultrasounds were performed in our centre over a period of six months. The scans were performed according to the Australian Society of Ultrasound and Medicine guidelines for the second trimester ultrasound examination.

Images were taken using ultrasound machines (General Electric Voluson 730, General Electric Voluson i, GE Medical Systems Austria; and Medison Accuvix V20, Medison Co. Ltd, Korea) with curvilinear abdominal transducers (GE – RAB 4-8 L and Medison – 3D 2-6 ET). One hundred normal cases were randomly selected and 10 images from each case were analysed (Figure 1). Each image was assessed with predefined criteria by four experienced reviewers (Table 1). One point was awarded for each criterion totalling a maximum score of 52.

The first 25 cases were designated as a preliminary study to train the reviewers on the criteria and the method of analysis. The remaining 75 cases were used to assess the benefits of a preliminary training on quantitative evaluation of sonographer performance in the fetal morphology ultrasound examination.

### Table 1: Criteria used for marking – each criterion was awarded 1 point totalling a maximum score of 52.

| Criteria | BPD/HC | Posterior Fossa | Lateral Ventricle | Nose/lips | Cord insertion | 4 chamber heart | LVOT | RVOT | Kidneys | Spine |
|----------|--------|-----------------|-------------------|-----------|----------------|-----------------|------|------|---------|-------|
| 1 | Cavum septum pellucidum visible | Dumbell-shaped cerebellum visible | Medial and lateral edge of posterior horn visible | Upper lip visible | Skin line on both sides of insertion visible | 4 chambers visible | Continuous IVS | Pulmonary trunk visible | Transverse view of 1st kidney visible | Dorsal spine visible |
| 2 | Thalami visible | Cisterna magna visible | Horizontal plane of image | Lower lip visible | Vessels seen within cord | Apex of heart visible | Left ventricle visible | Aorta in cross section visible | Transverse view of 2nd kidney | Sacrum visible |
| 3 | Falc visible | Nuchal fold visible | Symmetrical image | Both nostrils visible | Abdomen size greater than 2/3 | Crux visible | Fetal spine visible | Superior vena cava visible | Posterior kidney clear of spine | Alignment of vertebrae seen in lumbar spine |
| 4 | Bony detail clearly demonstrated | Cavum septum pellucidum visible | Choroid plexus visible | Image size greater than 1/3 | Pulmonary vein visible | Image size greater than 1/2 | Fetal spine in image | Renal pelves visible | Continuity of skin line |
| 5 | Horizontal plane of image | Horizontal plane of image | Image size greater than 1/2 | | Descending thoracic aorta visible | Image size greater than 1/2 | Abdomen size greater than 1/2 | Amniotic fluid seen beyond skin line |
| 6 | Symmetrical image | Symmetrical image | | | | Image size greater than 1/2 | | Region of interest greater than 1/2 |
| 7 | Image size greater than 2/3 | Image size greater than 2/3 | | | | | | | | |

**Figure 1:** Examples of the 10 ultrasound images analysed using the quantitative method – (A) Biparietal diameter/Head circumference; (B) Cerebellum; (C) Lateral ventricles; (D) Lips; (E) Cord insertion; (F) Four chamber heart; (G) Left ventricular outflow tract; (H) Right ventricular outflow tract; (I) Transverse kidneys; and (J) Lower sagittal spine.
training period and to compare the reproducibility of the results. The scores acquired by each reviewer were statistically analysed using Pearson’s and intra-class correlations to determine the reproducibility of the quantitative method. Further analysis of the raw data and statistical results was made to identify the inaccuracies of the quantitative method.

Results

Preliminary study

The correlation between reviewers was measured using the Pearson product-moment correlation coefficient. This coefficient measures the strength of dependence between two reviewers. The greatest amount of correlation was demonstrated between reviewers 1 and 4, yielding a Pearson’s correlation coefficient of 0.58. The results between reviewers 3 and 4 appeared to be independent of each other with a weak correlation coefficient of 0.05. The preliminary study results comparing all reviewers are illustrated in Table 2. A better appreciation of the correlation between reviewers is illustrated in the scatter plots with a line of best fit (Figure 2). The overall reviewer reproducibility was calculated using intra-class correlation which was reliant on the intra-reviewer and inter-reviewer variability. The preliminary intra-class correlation coefficient was 0.12.

Post training study

The best correlation was observed between reviewers 1 and 4 with a Pearson’s correlation coefficient of 0.71. The least amount of correlation was observed between reviewers 3 and 4 scoring 0.30. The correlation between reviewers in the post training study was tabulated (Table 3) and illustrated in Figure 3. The greatest correlation is depicted in the scatter plot between reviewer 1 and 4 with a strong positive linear relationship and a correlation coefficient closest to 1. An overall intra-class correlation coefficient of 0.24 was calculated, which demonstrated a significant difference between the intra-class correlation coefficient of the preliminary and the post training study. A two-fold increase in the correlation coefficient was observed after training the reviewers on the criteria. Improvement was seen following training however correlation remained low.

The raw data were reanalysed to identify possible problems in the criteria. The scores with a total difference of greater than 30% were extracted and the reviewer was interviewed to determine the issues that would assist in refining the model for future studies.

Discussion

The purpose of documenting images in the routine second trimester foetal anomaly scan is to provide evidence of developmental well-being. Therefore a high standard of imaging is critical in reflecting the quality of the scan. Poor image quality would imply that reassessment is necessary or an inadequate examination was performed.

Developing an effective method of assessing ultrasound
image quality is important in providing quality assurance. The
determination of image quality at present is subjective and
varies between reviewers. A more refined method is essential to
simplify and identify ultrasound quality.

This study provides a standardised method of interpreting
and quantifying the quality of the scan. The reproducibility of
the results was tested between four reviewers to determine the
reliability of the method before and after training. The images
selected were from the Australian Society of Ultrasound and
Medicine (ASUM) guidelines to demonstrate fetal anatomy in
the second trimester.

Many studies have attempted to quantify image quality
however this is the first to incorporate a training period to
minimise error. The preliminary study provides the reviewer
with clear expectations which standardises performance. This
component has added value to the study by providing consistent
and reliable results. An increase in the intra-class correlation
coefficient was seen between the preliminary and post training
study. This demonstrates the importance in training the
reviewers to achieve a more reproducible method in quantifying
the quality of an ultrasound image.

Medium to strong correlation is seen between reviewers
in the post training study demonstrating the potential for an
increase in the reproducibility of the results after training. Inter-
reviewer calculations revealed an overall moderate correlation.
Reviewer 3 consistently presented with the weakest correlation
value in the inter-reviewer correlation calculations; which, may
be the reason for the low intra-class correlation coefficient.

Reanalysis of the raw data identified multiple key factors
that negatively impacted the overall reviewer scores. Addressing
these factors would assist in refining the quantitative method.
Scores should be given according to image format; split screen
would only require a half-filled screen. In cases where there are
repeated images the best scoring view should be nominated.
Images should be scored irrespective of the fetal position. Pre-
selection of images may rectify the inadvertent omission of
views. Reviewer 3 consistently presented with the lowest score
and the largest difference compared to the other reviewers. The
consistency from reviewer 3 indicates that the outcome may
have been due to a lack of understanding and therefore a need
for additional training.

There are several other factors that may affect the overall
results. The study does not take into account that image quality
may be affected by fetal position or maternal characteristics, for
example maternal habitus. A poor quality image can be disguised
by a good score; each criterion is awarded one point regardless of
the importance value to the anatomy reviewed.

In order to improve on the quantitative study reviewer
variability must minimised. An instruction manual detailing the
correct interpretation of the criteria ought to be developed. The
manual would include a definition of each criterion to assist in
further training reviewers to improve reproducibility. A short
examination on the scoring process would also isolate any
difficulties the reviewers may have.

An addition of intra-reviewer correlation would have
determined the amount of variability in each reviewer. This
would strengthen the reliability of the results. These findings
demonstrate the need for further evaluation of the criteria and
that ongoing study is necessary to achieve a more precise model
of quantifying image quality.

---

Table 3: Post training study correlation coefficients between each reviewer.

| Reviewer | 1   | 2   | 3   | 4   |
|----------|-----|-----|-----|-----|
| 1        | 1.00|     |     |     |
| 2        | 0.54| 1.00|     |     |
| 3        | 0.36| 0.35| 1.00|     |
| 4        | 0.71| 0.61| 0.30| 1.00|

---

Figure 3: Scatter plot of the results illustrating the relationship between reviewer scores.
Conclusion
The quantitative method used in this study proved to be a suitable baseline model for assessing ultrasound images. This method reduces subjectivity and places a value on image quality. Image quality is more interpretable to laymen when presented as a value and may also be more appropriate for use in medico-legal cases.

A training period is beneficial in producing a more ideal study by increasing the reproducibility of the results; however an addition of an instruction manual on the method of assessment would be useful in improving the correlation between reviewers. This study should be viewed as the foundation of a more feasible and reliable method in quantifying image quality for the future.

Acknowledgements
We would like to thank the staff in the Department of Perinatal Ultrasound and Michael Eagleton for their assistance in this study. We would also like to acknowledge Andrew Martin for all the statistical work and support. Lastly, we are grateful to all the pregnant patients who underwent the ultrasound examination.

References
1. Salomon LJ, Winer N, Bernard JP, Ville Y. A score-based method for quality control of fetal images at routine second-trimester ultrasound examination. Prenat Diagn 2008; 28: 822–27.
2. Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. Feasibility and reproducibility of an image-scoring method for quality control of fetal biometry in the second trimester. Ultrasound Obstet Gynecol 2006; 27: 34–40.
3. Salomon LJ, Ville Y. The science and art of quality in obstetric ultrasound. Curr Opin Obstet Gynecol 2009; 21: 153–60.
4. AJUM. AJUM practice guideline for the performance of obstetric ultrasound examinations. J Ultrasound Med 2010; 29: 157–66.
5. ISUOG. Cardiac screening examination of the fetus: guidelines for performing the ‘basic’ and ‘extended basic’ cardiac scan. Ultrasound Obstet Gynecol 2006; 27: 107–13.
6. RCOG. Ultrasound screening. Available at http://www.rcog.org.uk/womens-health/clinical-guidance/ultrasound-screening. Accessed April 2010.
7. Angtuaco TL. Ultrasound imaging of fetal brain abnormalities. Ultrasound Q 2005; 21: 287–94.
8. Budorick NE, Pretorius DH, Nelson TR. Sonography of the fetal spine: technique, imaging findings, and clinical implications. AJR Am J Roentgenol 1995; 164: 421–28.
9. Chitty LS, Altman DG, Henderson A, Campbell S. Charts of fetal size: 2. Head measurements. Br J Obstet Gynaecol 1994; 101: 35–43.
10. Sairam S, Awadh AM, Cook K, Papageorghiou AT, Carvalho JS. Impact of audit of routine second-trimester cardiac images using a novel image-scoring method. Ultrasound Obstet Gynecol 2009; 33: 545–51.
11. Smijders RJ, Nicolaides KH. Fetal biometry at 14–40 weeks’ gestation. Ultrasound Obstet Gynecol 1994; 4: 34–48.
12. ASUM. Guidelines for the mid trimester obstetric scan. Available at http://www.asum.com.au/newsite/files/documents/policies/PS/D2_policy.pdf. Accessed April 2010.
13. Jaudi S, Tezenas Du Montcel S, Fries N, Nizard J, Halley Desfontaines V, Dommergues M. Online evaluation of fetal second-trimester four-chamber view images: a comparison of six evaluation methods. Ultrasound Obstet Gynecol 2011; 38: 185–90.
14. Sarris I, Ioannou C, Dighe M, Mitidieri A, Obertos M, Qingqing W, et al. Standardization of fetal ultrasound biometry measurements: improving the quality and consistency of measurements. Ultrasound Obstet Gynecol 2011; 38: 681–87.