

A score-based method for quality control of fetal images at routine second-trimester ultrasound examination

L. J. Salomon^{1,3}, N. Winer^{2,3}, J. P. Bernard^{1,3} and Y. Ville^{1,3*}

¹Department of Obstetrics and Gynecology, Université Paris-Ouest Versailles-St. Quentin, Centre Hospitalier Intercommunal Poissy-St. Germain, Poissy, France

²Department of Obstetrics and Gynecology, Centre Hospitalier Universitaire of Nantes, Nantes, France

³Société Française pour l'Amélioration des Pratiques Echographiques (SFAPE)

Objectives Our aim was to develop and evaluate the feasibility and reproducibility of score-based quality control for routine standardized fetal ultrasound images obtained in the second trimester of pregnancy.

Study Design In France, a minimum of three biometrical and six anatomical standardized ultrasound planes are to be produced with any mid-trimester scan. All anatomical standardized ultrasound images, routinely obtained by one trained operator at 20 to 24 weeks, were stored prospectively during a 1-year period. Twenty-five examinations containing these images were later randomly selected. These were then analyzed by two reviewers, according to predefined criteria agreed upon on the basis of established standards. This yielded a total score of up to 32 points. Feasibility, inter- and intra-reviewer reproducibility were analyzed.

Results Routine second-trimester ultrasound examinations numbering 1160 performed over a one year period by one trained sonographer unaware of the subsequent study at the time the images were recorded and stored in a database. Among the 150 images randomly selected and analyzed, adjusted kappa values were above 0.8 for 27 (84%) and 30 (94%) criteria, intra-class correlation coefficient was 0.86 (0.75; 0.96) and 0.98 (0.94; 1) and the mean difference (95% CI) in score was -0.44 (-3.0 ; 2.1) and -0.2 (-2 ; 1.6) for inter- and intra-reviewer comparisons respectively.

Conclusion A quality control policy based on image scoring is feasible and allows for good inter- and intra-reviewer reproducibility. Besides its potential for audit and quality control, this could also be useful during the training process. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: ultrasound; fetal; quality control; biometry; score

INTRODUCTION

Prenatal ultrasound (US) examination has become a screening tool offered to all pregnant women, at least once in the second trimester of pregnancy in most developed countries. In the United States, approximately 65% of pregnant women have at least one ultrasound examination (ACOG, 2004).

The main aims of prenatal ultrasound screening are to confirm fetal viability and well being as well as to look for malformation that could influence prenatal management. Sensitivity for prenatal detection of malformations by US ranges from 27.5 (Levi, 2002) to 96% (Sabbagha *et al.*, 1985; Queisser-Luft *et al.*, 1998). This wide variation may be due, at least in part, to significant differences in training and also due to the quality of the images obtained.

There is a role for training and quality-assurance programs in routine ultrasound examination. A standard method for performing ultrasound examination and images expected to be produced and examined have been agreed upon in various countries (RCOG, 2000;

AIUM, 2003; ACOG, 2004; Comité National Technique de l'Echographie de Dépistage Prénatal, 2005). All guidelines emphasize the need for documentation of US examinations and their importance for quality assurance. In France, at least nine images should be produced and recorded during any routine second-trimester ultrasound examination: three for standard ultrasound measurements and six for the assessment of key anatomical features (Comité National Technique de l'Echographie de Dépistage Prénatal, 2005) (Figure 1). The purpose of this study was to develop and evaluate the feasibility and reproducibility of a score-based quality control for the evaluation of these images required by the French standards.

MATERIALS AND METHODS

Standardized ultrasound images numbering 10 440, documenting 1160 routine second-trimester ultrasound examination performed over a 1-year period by one trained sonographer (J. P. Bernard), unaware of the subsequent study at the time of imaging, were recorded and stored in a database. All images were recorded at 20 to 24 weeks gestation with no time constraints using the same probe and ultrasound machine (3.5–5 MHz curvilinear abdominal transducer and 7 MHz vaginal

*Correspondence to: Y. Ville, Department of Obstetrics and Gynecology, Université Paris-Ouest Versailles-St. Quentin, Centre Hospitalier Intercommunal Poissy-St. Germain, 10 rue du Champ Gaillard, 78300 Poissy, France. E-mail: yville@wanadoo.fr

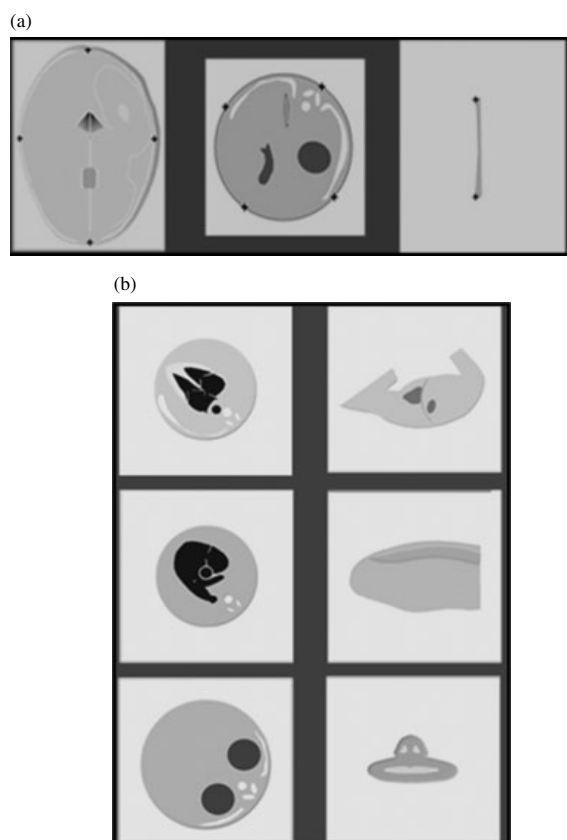


Figure 1—The nine minimum images that should be recorded at any standard ultrasound examination, chosen by the French National Ultrasound screening committee (Comité National Technique de l'Échographie de Dépistage Prénatal, 2005). The three biometrical images ((a) cephalic, abdominal and femur measurements) were not evaluated in this study but this has been reported elsewhere (Salomon *et al.*, 2006). The six anatomical standardized ultrasound planes were those evaluated herein (b)

transducer—General Electric Voluson 730 Expert-GE Medical System Europe-78 Buc France), with cine-loop facility. In France, 'trained operators' are obstetricians, radiologists or midwives who have been granted the National Diploma in obstetrical and gynaecological ultrasound examination. In this study, the trained operator performed more than 2000 ultrasound examinations per year and worked in agreement with the French guidelines on mid-trimester ultrasound examination. These guidelines include the requirement to produce and store at least the nine images illustrated in Figure 1. There was no Institutional Review Board (IRB) approval since this study did not modify routine prenatal care.

The three biometrical images (for head and abdominal circumference and femur length measurement respectively), for which the feasibility of a scoring method was previously demonstrated, were not included in this analysis (Salomon *et al.*, 2006). A total of 150 images comprised of 25 for each of the standardized anatomical planes (25 complete anatomical examinations) were

selected using random selection of 25 examinations out of the 1160 stored.

The 150 images were then projected on a wide screen during one session, to two reviewers (A and B) for an independent evaluation using an objective scoring method. These two reviewers were experienced ultrasound operators, who did not work together. Objective scoring was performed according to predefined criteria summarized in Table 1. These criteria were specific of each type of images and were agreed upon on the basis of established standards for fetal examination (RCOG, 2000; AIUM, 2003; ACOG, 2004; Comité National Technique de l'Échographie de Dépistage Prénatal, 2005). Each criterion scored one point when it was met, yielding a maximum score of up to 32 points for the six images of the anatomical ultrasound examination. Figure 2 illustrates ultrasound images meeting quality criteria. To assess intra-rater variability, ten complete examinations were scored twice by the first reviewer.

Statistical analysis

Normality of the distribution of scores was tested using Shapiro-Wilk's W -test. Inter-rater variability was tested as follows: the mean scores attributed by each reviewer were compared using a paired t -test. The percentage of images meeting the quality criteria was calculated. Adjusted kappa coefficients (Bennett *et al.*, 1954) for each criterion alone were computed in order to test for the reproducibility of each independent criterion. The intra-class correlation of scores given by the two reviewers was calculated and the difference in scoring was assessed by using the Bland–Altman method and plot (Bland and Altman, 1986). Intra-reviewer variability was assessed using the same method.

Statistical analyses were performed using Stata 9.2 for Windows (StataCorp LP, TX 77 845 USA), Statistica 6.0 (StatSoft, OK 74 104 USA) and Excel 2000 (Microsoft, Seattle, USA). For all tests, a value of $p < 0.05$ was considered statistically significant. Adjusted kappa values <0.6 , ≥ 0.6 and <0.8 , and ≥ 0.8 were taken as representing poor, moderate and good agreement, respectively (Cohen, 1960).

RESULTS

Scores given by both reviewers were normally distributed, as assessed by the Shapiro-Wilk's W -test ($p = 0.36$ and 0.08 for A and B, respectively). The mean (\pm SD) scores were not different between reviewers and were of $25.4 (\pm 2.8)$ and $25.8 (\pm 2.3)$ for reviewers A and B, respectively ($p > 0.05$).

The median (range) percentage of images meeting the quality criteria found to be present was 83% (48; 100). Adjusted kappa values between reviewers corresponding to each individual criterion were calculated for the 32 criteria assessed by the two reviewers. Values were below 0.6, between 0.6 and 0.8 and above 0.8 for 1 (3%), 4(13%) and 27 (84%) criteria respectively. All results are shown in Table 2.

Table 1—Criteria for score-based evaluation: each image was scored according to four to six criteria. Each criterion scored one point when it was met yielding a maximum score of 32 points for the six images of the anatomical ultrasound examination

Criteria	4-Chamber	Outflow tracts	Kidneys	Stomach/diaphragm	Spine	Face
1	4 chambers visible.	Pulmonary artery bifurcation visible	Circular view of the first kidney	Heart visible	Dorsal spine visible	Upper lip visible
2	Apex of the heart visible.	Ascending aorta visible	Circular view of the second kidney	Stomach visible	Sacrum visible	Two nostrils visible
3	Heart crux visible.	Right ventricle visible.	Posterior kidney clear from the spine acoustic shadow.	Spine nonvisible	Alignment of the vertebrae visible from the dorsal level to the sacrum	Two lip angles visible
4	One pulmonary vein visible	Pulmonary artery curling up the aorta.	Corticomedullary differentiation or pyelic cavity visible.	Diaphragmatic interface visible from back to front	Continuity of the skin line	—
5	Descending thoracic aorta visible.	—	—	Thigh and neck visible	Amniotic fluid visible beyond the skin	—
6	Region of interest occupying more than half of the total image size	Region of interest occupying more than half of the total image size	Region of interest occupying more than half of the total image size	Region of interest occupying more than half of the total image size	Region of interest occupying more than half of the total image size	Region of interest occupying more than half of the total image size

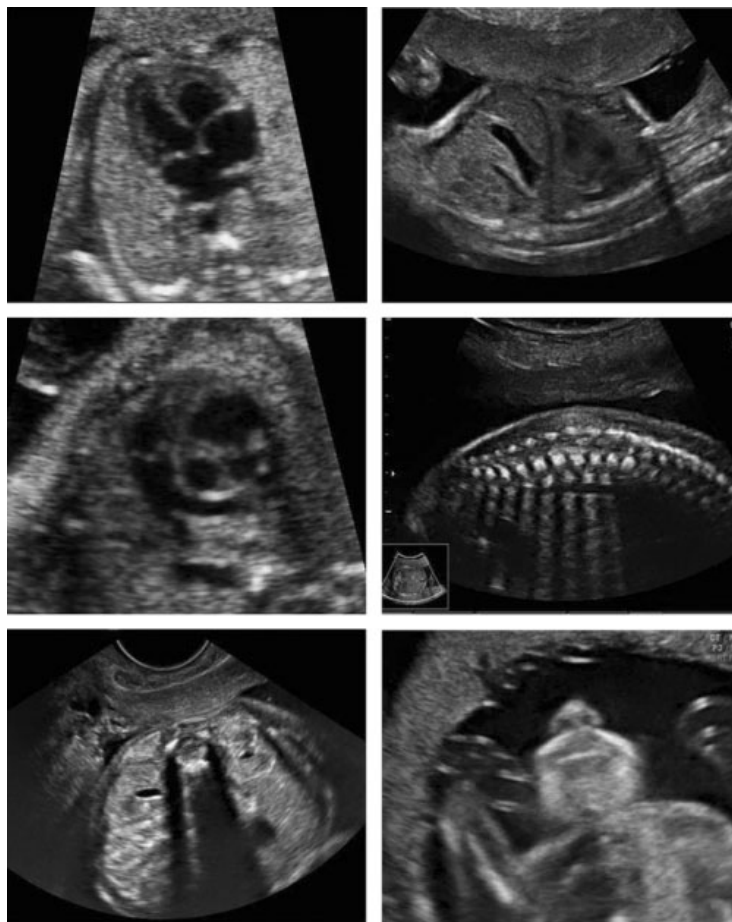


Figure 2—Example of anatomical images meeting our quality criteria

Copyright © 2008 John Wiley & Sons, Ltd.

Prenat Diagn 2008; **28**: 822–827.
DOI: 10.1002/pd

Table 2—Kappa coefficients for criteria

Kappa (κ)	Four-Chamber	Outflow tracts	Kidneys view	Stomach/diaphragm	Spine	Face
1	% of presence Four chambers visible : 68%	Pulmonary artery bifurcation visible: 84%	Circular view of the first kidney: 82%	Heart visible: 88%	Dorsal spine visible: 98%	Superior lip visible: 100%
Inter	1	1	0.84	1	0.92	1
Intra	1	0.8	1	1	1	1
2	Apex of the heart visible: 84%	Ascending aorta visible: 100%	Circular view of the second kidney: 72%	Stomach visible: 84%	Sacrum visible: 74%	Two nostrils visible: 78%
Inter	0.84	1	0.92	1	0.92	0.92
Intra	0.8	0.6	0.8	1	0.8	1
3	Heart crux visible: 60%	Right ventricle visible: 66%	Posterior kidney clear from the spine acoustic shadow: 92%	Smine nonvisible: 62%	Alignment of vertebrae visible from the dorsal level to the sacrum: 96%	Two lip angles visible: 72%
Inter	0.84	0.6	0.84	0.76	1	1
Intra	1	1	1	0.6	1	1
4	One pulmonary vein visible: 78%	Pulmonary artery curling up the aorta: 86%	Corticomedullary differentiation or pyelic cavity visible: 80%	Diaphragmatic interface visible from back to front : 80%	Continuity of the skin line: 56%	—
Inter	0.92	0.92	0.52	0.68	1	—
Intra	0.6	1	0.8	1	1	—
5	Descending thoracic aorta visible : 88%	—	—	Thigh and neck visible: 48%	Amniotic fluid visible beyond the skin: 66%	—
Inter	0.84	—	—	0.84	0.6	—
Intra	1	—	—	0.8	1	—
6	Region of interest occupying more than half of the total image size: 88%	Region of interest occupying more than half of the total image size: 88%	Region of interest occupying more than half of the total image size: 84%	Region of interest occupying more than half of the total image size: 80%	Region of interest occupying more than half of the total image size: 96%	Region of interest occupying more than half of the total image size: 88%
Inter	1	1	0.84	1	1	1
Intra	1	1	1	1	1	1

Intra-class correlation coefficient was 0.86 (0.75; 0.96) and the mean difference (95% CI) in score was -0.44 (-3.0 ; 2.1). The stability of the mean difference is illustrated in the Bland and Altman plot (Figure 3).

There was no difference in mean scores attributed by the same reviewer during the two scoring sessions (24.1 ± 4.2 and 23.9 ± 4.3 for first and second scoring respectively, $p > 0.05$). Adjusted kappa values were below 0.6, between 0.6 and 0.8 and above 0.8 for 0(0%), 2(6%) and 30 (94%) criteria respectively. Intra-class correlation coefficient was 0.98 (0.94; 1) and the mean difference (95% CI) in score was -0.2 (-2 ; 1.6).

DISCUSSION

Although the need for documentation of US examinations in pregnancy is recommended by national colleges or ultrasound societies (RCOG, 2000; AIUM, 2003; ACOG, 2004; Comité National Technique de

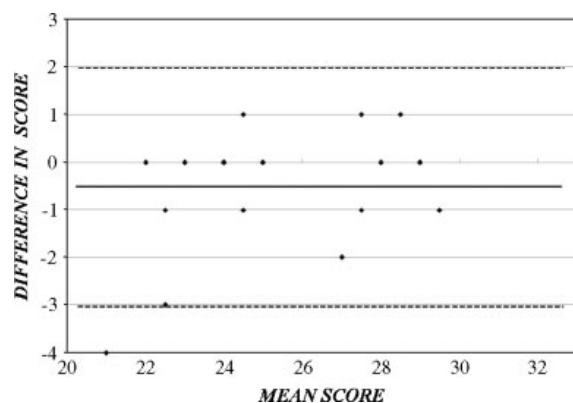


Figure 3—Bland and Altman plot for the agreement in scoring for reviewers A and B

l'Echographie de Dépistage Prénatal, 2005), the relevant information that can be extracted from these images

for quality control has never been evaluated. Physicians are held responsible for the quality of the documentation of examinations as well as of the quality control and safety of both the environment and the procedure. Quality assessment is facilitated by storage of appropriate digital images in the medical record or in another database. This study demonstrates that reliable quality control can be performed based on such images.

Quality control studies in prenatal sonography have long been based on detection rates of fetal abnormalities (Nelson *et al.*, 1993). Although this is an important issue, congenital anomalies are encountered in no more than 2% of all pregnancies (Anderson *et al.*, 1995) and a prevalence bias precludes wide usage of this approach. Moreover, these processes have not prevented a surge in malpractice law suits for missed fetal anomalies, which have become the most common type of litigation involving ultrasound (Sanders, 1998). Training and certification programs are likely to improve quality of ultrasound screening and simple and reproducible quality control processes are needed (AIUM, 2003). Image scoring in prenatal ultrasound was first introduced as the corner stone of quality control of nuchal translucency measurements at 11 to 14 weeks of gestation (Herman *et al.*, 1998; Snijders *et al.*, 1998, 2002; Dudley and Chapman, 2002; Fries *et al.*, 2007) and was also proposed for controlling biometric images (Salomon *et al.*, 2006).

The nine images chosen include all measurements as well as essential elements of the fetal anatomy, as defined by the American Colleges of Obstetrics and of Radiology (Radiology, 2003; ACOG, 2004). Quality control of these images is, therefore, likely to reflect that of the fetal anatomical survey (Abuhamad *et al.*, 2004). Although this correlation remains speculative, poor image documentation precludes the possibility for reinterpretation and could support allegations that an incomplete or inadequate study has been performed (ACOG, 2004).

Our study has several limitations. The process of image review requires significant human resources. Extensive qualitative analysis would be time-consuming and expensive and the review of a small sample of images may provide a time-limited or image-limited quality assessment. However, our study demonstrates that quality assessment of ultrasound examination is feasible and reproducible using a simple score. We did not aim at measuring the true quality of an examination or of an operator. Indeed, ultrasound is a dynamic examination and its performance in a given patient can hardly be evaluated based on a sample of images. Moreover, quality might be affected by fetal position or maternal characteristics. Nevertheless, our scoring system allows sonographers to assess the overall quality of their practice.

Qualitative review also requires the definition of criteria whose choice may seem arbitrary and may also introduce a bias in quality assessment when a systematic mistake for one criterion is masked by a good overall score. Although one could argue that the chosen criteria cannot be strictly considered as objective, reviewers easily agree on most criteria when scoring the images but some criteria performed worse than

others. There was perfect agreement for inter- and intra-reviewers' comparisons when assessing the four-chamber view of the fetal heart. However, visualization of corticomedullary differentiation or visualization of a pyelic cavity had the poorest performance, with a kappa value of 0.52. Although we demonstrated good reliability and reproducibility for this criteria-based score, the results might have been different by using a different system or by using different rules to evaluate agreement between reviewers. One might also decide to give different weight to different criteria in order to emphasize a particular aspect of quality. However, our results support the chosen criteria. Although the sonographer was unaware of the subsequent study and chosen criteria, these were found to be present in most images. All but one criteria (Table 2) were present in more than 50% of images with a median incidence of 83%.

The contrast between increasing expectations in antenatal care and limited health care resources reinforces the necessity to ensure that screening procedures are clinically and economically effective. Because of the increasing impact of legal aspects in ultrasound practice, it has also become necessary for sonologists to develop systems that could help in certifying and assessing the overall quality of their practice. Besides its potential for audit and quality control, this image-scoring method could also be useful during the process of training through performance evaluation as well as correcting systematic errors of both trainees and trained operators.

REFERENCES

- Abuhamad AZ, Benacerraf BR, Woletz P, Burke BL. 2004. The accreditation of ultrasound practices: impact on compliance with minimum performance guidelines. *J Ultrasound Med* **23**(8): 1023–1029.
- ACOG. 2004. ACOG Practice Bulletin No. 58. Ultrasonography in pregnancy. *Obstet Gynecol* **104**(6): 1449–1458.
- AIUM. 2003. AIUM Practice Guideline for the performance of an antepartum obstetric ultrasound examination. *J Ultrasound Med* **22**(10): 1116–1125.
- Anderson N, Boswell O, Duff G. 1995. Prenatal sonography for the detection of fetal anomalies: results of a prospective study and comparison with prior series. *AJR Am J Roentgenol* **165**(4): 943–950.
- Bennett E, Albert R, Goldstein A. 1954. Communications through limited response questioning. *Public Opin Q* **18**: 303–308.
- Bland JM, Altman DG. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**(8476): 307–310.
- Cohen J. 1960. A coefficient of agreement for nominal scales. *Educ Psychol Meas* **20**: 37–46.
- Comité National Technique de l'Échographie de Dépistage Prénatal. 2005. Rapport du Comité National Technique de l'Échographie de Dépistage Prénatal, 81.
- Dudley NJ, Chapman E. 2002. The importance of quality management in fetal measurement. *Ultrasound Obstet Gynecol* **19**(2): 190–196.
- Fries N, Althuser M, Fontanges M *et al.* 2007. Quality control of an image-scoring method for nuchal translucency ultrasonography. *Am J Obstet Gynecol* **196**(3): 272e1–272e5.
- Herman A, Maymon R, Dreazen E, Caspi E, Bukovsky I, Weinraub Z. 1998. Nuchal translucency audit: a novel image-scoring method. *Ultrasound Obstet Gynecol* **12**(6): 398–403.
- Levi S. 2002. Ultrasound in prenatal diagnosis: polemics around routine ultrasound screening for second trimester fetal malformations. *Prenat Diagn* **22**(4): 285–295.

- Nelson NL, Filly RA, Goldstein RB, Callen PW. 1993. The AIUM/ACR antepartum obstetrical sonographic guidelines: expectations for detection of anomalies. *J Ultrasound Med* **12**(4): 189–196.
- Queisser-Luft A, Stopfkuchen H, Stolz G, Schlaefer K, Merz E. 1998. Prenatal diagnosis of major malformations: quality control of routine ultrasound examinations based on a five-year study of 20,248 newborn fetuses and infants. *Prenat Diagn* **18**(6): 567–576.
- Radiology ACO. 2003. ACR practice guideline for the performance of antepartum obstetrical ultrasound. *ACR Practice Guidelines and Technical Standards*. ACR: Philadelphia, PA.
- RCOG. 2000. *Ultrasound Screening for Fetal Abnormalities*. RCOG: London.
- Sabbagha RE, Sheikh Z, Tamura RK, *et al.* 1985. Predictive value, sensitivity, and specificity of ultrasonic targeted imaging for fetal anomalies in gravid women at high risk for birth defects. *Am J Obstet Gynecol* **152**(7 Pt 1): 822–827.
- Salomon LJ, Bernard JP, Duyme M, Doris B, Mas N, Ville Y. 2006. Feasibility and reproducibility of an image scoring method for quality control of fetal biometry in the second trimester. *Ultrasound Obstet Gynecol* **27**(1): 34–40.
- Sanders RC. 1998. Legal problems related to obstetrical ultrasound. *Ann N Y Acad Sci* **847**: 220–227.
- Snijders RJ, Noble P, Sebire N, Souka A, Nicolaides KH, Fetal Medicine Foundation First Trimester Screening Group. 1998. UK multicentre project on assessment of risk of trisomy 21 by maternal age and fetal nuchal-translucency thickness at 10–14 weeks of gestation. *Lancet* **352**(9125): 343–346.
- Snijders RJ, Thom EA, Zachary JM, *et al.* 2002. First-trimester trisomy screening: nuchal translucency measurement training and quality assurance to correct and unify technique. *Ultrasound Obstet Gynecol* **19**(4): 353–359.