

Nuchal translucency audit: a novel image-scoring method

A. Herman, R. Maymon, E. Dreazen, E. Caspi*, I. Bukovsky and Z. Weinraub

Department of Obstetrics and Gynecology, Assaf Harofeh Medical Center, Zerifin; *The Tarnesby Tarnowsky Chair for Family Planning and Fertility Regulation, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

Key words: NUCHAL TRANSLUCENCY, ULTRASOUND SCREENING, ONGOING AUDIT, IMAGE-SCORING METHOD

ABSTRACT

Objective The aim of this study was to evaluate the feasibility and reproducibility of a novel image-scoring method of first-trimester nuchal translucency measurement as an objective tool of ongoing audit and training.

Design This was an independent evaluation of nuchal translucency images by three separate reviewers unaware of the examiner.

Subjects There were 105 consecutive singleton pregnancies undergoing first-trimester screening.

Methods Each image was scored according to the following criteria: section (oblique, 0; mid-sagittal, 2), caliper placing (misplaced, 0; proper, 2), skin line (nuchal only, 0; nuchal and back, 2), image size (unsatisfactory, 0; satisfactory, 1), amnion (not visualized, 0; visualized, 1) and head position (flexion/hyperextension, 0; straight, 1). The final score was categorized into one of four quality groups: excellent (8–9), reasonable (4–7), intermediate (2–3), unacceptable (0–1).

Results The distributions of the four quality groups were similar between the three reviewers: 11.4% were classified as excellent, 57.1% as reasonable, 25.7% as intermediate and 5.7% as unacceptable. Inter-reviewer agreement showed identical classification, by each pair of reviewers, from 65.7% to 74.3%, and partial agreement to neighboring quality groups from 25.7% to 34.3% of the cases. In none of the cases did the reviewers differ in categorizing cases to remarkably different quality groups. Application of the auditing method to the examiners showed similar distribution to the various quality groups and similar mean final score of 4.69 (0.39, SE), 4.54 (0.15, SE) and 4.65 (0.15, SE).

Conclusions The described image-scoring method represents a new approach towards the evaluation of ultrasound performance as a whole and nuchal translucency measurement in particular. It may be employed by every center in

an independent manner with minimal resources and regardless of the method of risk assessment. More studies will be needed to determine the standards required from the examiners and to elucidate the contribution of the proposed auditing method to the examination's quality and the process of training.

INTRODUCTION

First-trimester nuchal translucency measurement has turned from a stimulating project unique to cases undergoing invasive procedure^{1–3} into an established screening program, applied to the general population in order to detect fetal aneuploidy, especially Down's syndrome^{4–7}. Accordingly, a study group of the Royal College of Obstetricians and Gynaecologists (RCOG) approved it as 'an acceptable method' and focused on the need for 'external systems of quality assurance and ongoing audit'⁸.

The Fetal Medicine Foundation⁶, which promoted the concept of first-trimester screening, issued technical guidelines for the examination and developed a special process of accreditation. Regardless of the method of risk assessment, either by a combination of maternal background risk and a likelihood ratio derived from the examination^{4,9}, or by using a cut-off value^{3,5,7,10}, the need for a meticulous measurement is obvious. Therefore, a process of quality assurance and ongoing audit, in centers implementing the program, seems to be mandatory. Although uniformity of the generated image and the measurement has become universal^{3,5–7,10} and the need for special qualification and training programs has become obvious^{6,11,12}, the issue of ongoing audit has not yet gained sufficient attention. The sole systematic method of ongoing audit, utilized by the Fetal Medicine Foundation, includes a computerized comparison of medians of measurements and an occasional evaluation of the images. Others, including the RCOG,

Correspondence: Professor A. Herman, Department of Obstetrics and Gynecology, Assaf Harofeh Medical Center, Zerifin, 70300, Israel

have not mentioned or proposed any specific method for ongoing audit.

The aim of the present study was to evaluate the feasibility and reproducibility of a novel method of scoring first-trimester nuchal translucency images, as a systematic objective tool of ongoing audit and training.

MATERIALS AND METHODS

The program of first-trimester screening is operated at our hospital by three qualified examiners, each having experience of more than 200 cases. Mostly, the examinations are performed with a 3.5-MHz curvilinear abdominal transducer (Advanced Technology Laboratories, HDI 3000, Seattle, WA, USA). The vaginal approach is reserved for those cases where difficulty is experienced in obtaining a proper image; this occurs in 5–10% of cases. The ultrasound machine allows a play-back facility, which is often used to obtain the best image. The images are generated according to the guidelines published by the Fetal Medicine Foundation⁶. These include a mid-sagittal section, sufficient magnification of the image to at least three-quarters of the screen, differentiation between fetal skin and amnion, and nuchal translucency measurement by placing the calipers on the white line so that the maximal translucent area is measured. Each measurement is video-printed twice and copies of the images that are used to assess the risk for trisomy 21 are kept in the patients' files.

The study included nuchal translucency images of 105 consecutive singleton pregnancies, each of the three examiners contributing 35 cases. The images, used for assessing the risk, were collected from each case and were mixed to avoid the identification of either the patient or the examiner. The study design comprised 105 separate and independent evaluations of the images by each of the reviewers assigned as reviewer A, B and C.

Each image was reviewed and scored according to the criteria shown in Figure 1. Three criteria (section, caliper placement and skin line) were assigned as major, and were each given a maximum score of 2 points, provided that the requirements were met. The other three criteria (image size, amniotic membrane and head position) were considered as minor, each providing 0 or 1 point. The sum of the points obtained from each criterion yielded the final score.

In order to avoid wide dispersion of the results, final scores were grouped according to quality groups (Table 1). A final score of 8 or 9 was considered an excellent examination, requiring maximal scores in almost all of the criteria. A reasonable examination could be considered as one with a score of 4–7, requiring at least two major criteria, or one major criterion and two minor ones. The intermediate group comprised cases with a score of 2 or 3, which required at least one major or two minor criteria. Cases with a score of 0 or 1 were categorized into the unacceptable group.

The reproducibility of the scoring method was assessed by comparing the distribution of the quality groups amongst the reviewers and by comparing the mean final score assigned by each reviewer. Also, inter-reviewer agree-

ment was assessed by analyzing mutual categorization to the various quality groups. The audit's scoring method was applied to each examiner by evaluating the distribution of categorization to quality groups and the mean final score of the actual examinations performed. The χ^2 test was used to analyze the distribution of the reviewers to quality groups, and analysis of variance was used to compare means of final score between the reviewers and the examiners.

RESULTS

Inter-reviewer comparison of the distribution of the cases according to various quality groups showed similar categorization by the reviewers (Table 2). The three reviewers classified 5.7–7.6% of the images as unacceptable, 21.9–30.5% of the images as intermediate, 51.4–61% as reasonable and 10.5–14.3% of the images as excellent. The differences between the reviewers were statistically not significant (χ^2 test, 11 degrees of freedom). Most of the images (71.5%, 66.7% and 63.8%) were classified as reasonable or excellent. Nevertheless, a substantial portion (36.1%, 33.3% and 28.6%) of the cases were classified as unacceptable or intermediate. This reflects a strict and unbiased attitude of the reviewers towards their own work. The mean final scores, assigned by each reviewer, were almost identical, being between 4.62 and 4.68, located in the reasonable quality group. The slight differences between the reviewers were not statistically significant (analysis of variance). Altogether, the total audit of all images, with the average final score of every case calculated, demonstrated more than 50% as reasonable, approximately 25% as intermediate, 10% as excellent and 5% as unacceptable.

Table 3 presents the agreement between reviewers concerning categorization to the different quality groups. Complete agreement (identical categorization by each pair of reviewers) was found in 65.7% of the cases between reviewers A and B, 68.6% between reviewers A and C and in 74.3% of the cases between reviewers B and C. Partial agreement (categorization of the reviewers to neighboring quality groups) was found in 25.7–34.3% of the cases. In none of the cases did the reviewers differ in categorizing cases to non-neighboring groups (excellent and intermediate groups or reasonable and unacceptable groups). Detailed inter-reviewer agreements, within the framework of the various quality groups, are specified in Figure 2. Every evaluation, between reviewers A and B, reviewers A and C and reviewers B and C, showed similar findings. The reviewers categorized most of the cases identically and partial agreement was distributed similarly on both sides of the central agreement line, indicating that none of the reviewers skewed in scoring the images, being more strict or permissive than his colleagues.

An example of an application of the suggested audit, applied to the examiners, is specified in Table 4. Each image was assigned a final score, by calculation of the mean value of the reviewers' scores. This audit, of the images assessed by each examiner, demonstrated a similar quality of performance amongst the three examiners, expressed as a non-significant difference between the mean













Criteria	Finding	Score	Finding	Score
Section	Oblique 	0	Mid-sagittal 	2
Caliper placement	Misplaced 	0	Proper placement 	2
Skin line	Nuchal only 	0	Nuchal and back 	2
Image size	Unsatisfactory 	0	Satisfactory 	1
Amnion	Not demonstrated 	0	Demonstrated 	1
Head position	Flexion/hyperextension 	0	Straight 	1

Figure 1 Nuchal translucency scoring according to the various criteria. Section, caliper placement and skin line constitute major criteria (score 0 or 2). Image size, amnion and head position constitute minor criteria (score 0 or 1)

final scores and similar distribution of the various quality groups. Moreover, the audit demonstrated similar standards of quality utilized by the examiners and that higher quality may be achieved in our center by broad improvement of the performance of all the examiners.

DISCUSSION

The program of first-trimester ultrasound screening presents a new challenge. Sonography, a modality based on subjective impression and semi-accurate measurements, is

turning into a diagnostic tool requiring particular imaging and a highly accurate measurement technique. Inter- and intraobserver variability, as well as specific conditions in each examination, affect the reproducibility of nuchal translucency measurements. Nevertheless, the examiner is required to produce a result that resembles laboratory tests, allocating patients to risk groups. Therefore, as in any laboratory study, those implementing the program should take upon themselves the task of an ongoing audit to adhere to the required standards and to avoid abuse of the method.

The issue of quality assurance in obstetric ultrasound examination usually focuses on the process of qualification and accreditation¹³. Furthermore, the quality of obstetric ultrasound examination is generally assessed by surveys concerning fetal malformation detection¹⁴⁻¹⁷. However, to the best of our knowledge, none of the previous studies

have suggested a method of assessing the quality of obstetric ultrasound examination by evaluating the images obtained. Thus, besides suggesting a novel method for nuchal translucency ongoing audit, the present study is the first study to utilize image review to assess the quality of the examinations.

The criteria and their corresponding scores deserve further explanation. Four of them, i.e. mid-sagittal section, caliper placement, image size and amnion visualization, were mentioned by the Fetal Medicine Foundation^{2,6}. More weight, however, should be given to the mid-sagittal section, which constitutes the key for proper examination, and to caliper placement, which was found to be a cardinal factor, explaining most of the inter- and intraobserver

Table 1 Categorization of the images into various quality groups

Score	Quality group	Minimum criteria required
8-9	excellent	3 major + 2 minor
4-7	reasonable	2 major or 1 major + 2 minor
2-3	intermediate	1 major or 2 minor
0-1	unacceptable	less than above

Table 2 Inter-reviewer comparison of the distribution* to quality groups and mean final score. Total is based on average final score of each reviewer

Quality group	Reviewer A (n = 105)	Reviewer B (n = 105)	Reviewer C (n = 105)	Total (n = 105)
Unacceptable	8 (7.6%)	7 (6.7%)	6 (5.7%)	6 (5.7%)
Intermediate	27 (25.7%)	23 (21.9%)	32 (30.5%)	27 (25.7%)
Reasonable	55 (52.4%)	64 (61.0%)	54 (51.4%)	60 (57.1%)
Excellent	15 (14.3%)	11 (10.5%)	13 (12.4%)	12 (11.4%)
Mean (± SE) score	4.68 (0.22) [†]	4.62 (0.21) [†]	4.65 (0.21) [†]	4.65 (0.21)

*Not significant, χ^2 test (11 degrees of freedom); [†]not significant, analysis of variance

Table 3 Inter-reviewer agreement of categorization to quality groups

Reviewers compared	Complete agreement*	Partial agreement [†]	Disagreement [‡]
A and B (n = 105)	69 (65.7%)	36 (34.3%)	0
A and C (n = 105)	72 (68.6%)	33 (31.4%)	0
B and C (n = 105)	78 (74.3%)	27 (25.7%)	0
Combined (n = 315)	219 (69.5%)	96 (30.4%)	0

*, Identical categorization to the same quality group; [†], close categorization to neighboring quality groups; [‡], categorization to far quality groups (e.g. unacceptable and reasonable)

Table 4 Audit of nuchal translucency images according to examiners by categorization to quality groups and to mean final score

Quality group	Examiner A (n = 35)	Examiner B (n = 35)	Examiner C (n = 35)
Unacceptable	3 (8.6%)	1 (2.9%)	2 (5.7%)
Intermediate	9 (25.7%)	9 (25.7%)	9 (25.7%)
Reasonable	18 (51.4%)	23 (65.7%)	19 (54.3%)
Excellent	5 (14.3%)	2 (5.7%)	5 (14.3%)
Mean (± SE) score	4.69 (0.39)*	4.54 (0.15)*	4.65 (0.47)*

*Not significant, analysis of variance

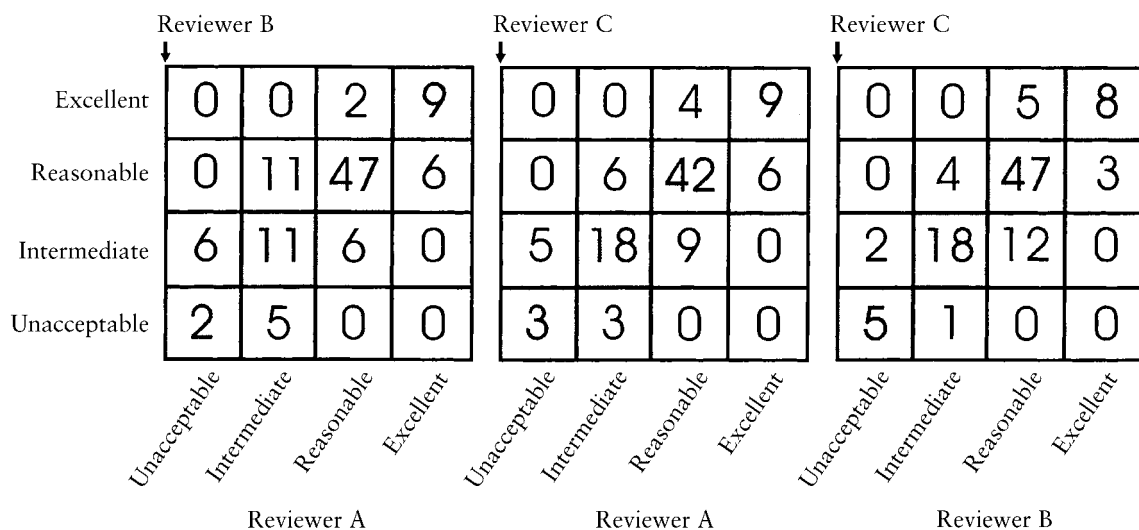


Figure 2 Inter-reviewer agreement within the framework of the unacceptable, intermediate, reasonable and excellent quality groups. The shaded areas represent complete agreement between the reviewers

variability¹⁸. Although differentiation between fetal skin and the amnion is mandatory during the dynamic examination, the demonstration of the amnion on the video-printed image is less important. Therefore, a lower score was assigned to amnion visualization and image size. We added two criteria that were not mentioned by the Fetal Medicine Foundation. The first was the skin line, whose presence at the nuchal region is mandatory and deserves no score. On the other hand, visualization of a continuous skin line along the fetal nuchal region and back remarkably improves the examination quality, turning it into an excellent one. The second criterion is fetal head position, which may affect nuchal translucency thickness, in cases of extreme flexion or hyperextension. Before the utilization of the present method, we conducted a pilot study in which the major criteria were divided into three categories so that images with borderline quality yielded 1 point. We noted several unsatisfactory images that attained an unjustified reasonable final score by gathering points. Therefore, major criteria included two categories only: 2 points, if the demands for high standards were met, or zero. The importance of image size and head position should not be underestimated, as these constitute part of the scoring method. Nevertheless, we believe that their weight is less important when evaluating the quality of image generation, and therefore these were scored with 1 point only. Final construction of the scoring system included three major criteria providing 2 points each, and three minor criteria providing 1 point each.

The Fetal Medicine Foundation is currently utilizing a computerized method that compares measurements, expressed in multiples of medians, with their own results. Although this method constitutes a breakthrough concerning this subject, it is of limited potential, since reasonable medians do not guarantee proper imaging. Furthermore, this method cannot differentiate specific errors, needing to be corrected. Moreover, their method is exclusive to those using their software and may not be applied to centers using a different approach to imaging or measurement. Our proposed image-scoring method overcomes most of these limitations:

- (1) It may be applied to any method of risk assessment;
- (2) Centers using a different method of imaging or measurement may accommodate it accordingly;
- (3) Minimal resources are needed;
- (4) It may be employed by every center in an independent manner;
- (5) There is a possibility for pointing out errors related to specific criteria.

However, similar to any auditing method, our scoring system is not free from limitations. Although the results may be computerized, the process of image review is based on human work. Attention should be paid to avoid bias and the audit should be preferably carried out by external

systems. The computerized audit and our scoring system should be regarded as complementary methods of ongoing audit rather than competitive.

Our image-scoring method also seems to be of potential value during the process of training. Braithwaite and colleagues¹¹ found that 80 abdominal and 100 vaginal examinations were needed before a trainee could be considered as trained. Only after performing these numbers of scans did the trainees achieve a desired repeatability coefficient in repeated blind measurements. In another study¹⁹ we showed that image magnification does not contribute to the repeatability of caliper placement. On the other hand, we found that repeated blind measurements on normal-sized images, and the corresponding magnified still images, may be utilized as an excellent method of training for proper placement of the calipers. The present scoring method seems to be a useful tool for training, enabling the evaluation of a broad spectrum of performance of the trainee, identifying specific errors and evaluating progress. Incorporation of the scoring method into the training program may reduce the number of examinations needed.

The fact that almost one-third of the images were classified as intermediate or unacceptable is troublesome. This highlights occasional difficulties in generating good images as well as the need for a more meticulous attitude. We assume that those findings represent the standard of performance by experienced examiners who are unaware of a process of image evaluation. It would be of interest to evaluate the results after the implementation of such ongoing audit.

In summary, the described image-scoring method proposes a new approach for the assessment of the quality of ultrasound examination as a whole and nuchal translucency measurement in particular. The feasibility of the method was demonstrated in this study and reasonable inter-reviewer agreements were obtained. More studies will be needed to confirm the applicability and weight of each criterion and to determine the standards required from the examiners. Further ongoing experience is required to elucidate the contribution of this novel ongoing audit to the quality of examinations and training.

REFERENCES

1. Szabo J, Gellen J. Nuchal fluid accumulation in trisomy-21 detected by vaginosonography in first trimester. *Lancet* 1990; 336:1133
2. Cullen MT, Gabrielli S, Green JJ, Rizzo N, Mahoney MJ, Salafia C, Bovicelli L, Hobbins JC. Diagnosis and significance of cystic hygroma in the first trimester. *Prenat Diagn* 1990;10: 643-51
3. Nicolaides KH, Azar G, Byrne D, Mansur C, Marks K. Fetal nuchal translucency: ultrasound screening for chromosomal defects in first trimester of pregnancy. *Br Med J* 1992;304: 867-9
4. Pandya PP, Snijders RJM, Johnson SP, Brizot ML, Nicolaides KH. Screening for fetal trisomies by maternal age and fetal nuchal translucency thickness at 10 to 14 weeks of gestation. *Br J Obstet Gynaecol* 1995;102:957-62

5. Hafner E, Schuchter K, Philipp K. Screening for chromosomal abnormalities in an unselected population by fetal nuchal translucency. *Ultrasound Obstet Gynecol* 1995;6:330-3
6. Snijders RJM, Johnson S, Sebire NJ, Noble PL, Nicolaides KH. First-trimester ultrasound screening for chromosomal defects. *Ultrasound Obstet Gynecol* 1996;7:216-26
7. Taipale P, Hiilesmaa V, Salonen R, Ylöstavo P. Increased nuchal translucency as a marker for fetal chromosomal defects. *N Engl J Med* 1997;337:1654-8
8. Recommendations arising from the 32nd study group: screening for Down's syndrome in the first trimester. In Grudzinkas JG, Ward RHT, eds. *Screening for Down's syndrome in the First Trimester*. London: Royal College of Obstetricians and Gynaecologists, 1977;353-6
9. Biagiotti R, Periti E, Brizzi L, Vanzi E, Cariati E. Comparison between two methods of standardization for gestational age differences in fetal nuchal translucency measurement in first-trimester screening for trisomy 21. *Ultrasound Obstet Gynecol* 1997;9:248-52
10. Pajkrt E, Bilardo CM, Van Lith JNM, Mol BWJ, Bleker OP. Nuchal translucency measurement in normal fetuses. *Obstet Gynecol* 1995;86:994-7
11. Braithwaite JM, Kadir RA, Pepera TA, Morris RW, Thompson PJ, Economides DL. Nuchal translucency measurement: training of potential examiners. *Ultrasound Obstet Gynecol* 1996;8:192-5
12. Pandya PP, Goldberg H, Walton B, Riddle A, Shelley S, Snijders RJM, Nicolaides KH. The implementation of first-trimester scanning at 10-13 weeks' gestation and the measurement of fetal nuchal translucency thickness in two maternity units. *Ultrasound Obstet Gynecol* 1995;5:20-5
13. Papp Z. Quality assurance in obstetric and gynecological ultrasound in Hungary. *Ultrasound Obstet Gynecol* 1996;7:305-6
14. Eik-Nes SH, Okland O, Aure JC, Ulstein M. Ultrasound screening in pregnancy: a randomized controlled trial. *Lancet* 1984;1:1347
15. Saari-Kemppainen A, Karjalainen O, Ylostalo P, Heinonen OP. Ultrasound screening and perinatal mortality: controlled trial of systematic one-stage screening in pregnancy. *Lancet* 1990;336:387-91
16. Levi S, Hyjazi Y, Schaaps J-P, Defoort P, Coulon R, Buekens P. Sensitivity and specificity of routine antenatal screening for congenital anomalies by ultrasound: the Belgian Multicentric Study. *Ultrasound Obstet Gynecol* 1991;1:102-10
17. Bernaschek G, Stuempflen I, Deutinger J. The influence of the experience of the investigator on the rate of sonographic diagnosis of fetal malformations in Vienna. *Prenat Diagn* 1996;16:807-11
18. Pandya PP, Altman DG, Brizot ML, Pettersen H, Nicolaides KH. Repeatability of measurement of fetal nuchal translucency thickness. *Ultrasound Obstet Gynecol* 1995;5:334-7
19. Herman A, Maymon R, Dreazen E, Caspi E, Bukovsky I, Weinraub Z. Image magnification does not contribute to the repeatability of caliper placement in measuring nuchal translucency thickness. *Ultrasound Obstet Gynecol* 1998;11:266-70