*Ultrasound Obstet Gynecol* 2014
Published online in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.13404

# Interobserver agreement in assigning IOTA color score to adnexal masses using three-dimensional volumes or digital videoclips: potential implications for training

L. PINEDA*, E. SALCEDO†, C. VILHENA‡, L. JUEZ* and J. L. ALCÁZAR*

**Department of Obstetrics and Gynecology, Clínica Universidad de Navarra, University of Navarra, Pamplona, Spain; †Department of Obstetrics and Gynecology, Clínica del Prado, University of Antioquía, Medellín, Colombia; ‡Department of Obstetrics and Gynecology, Hospital Garcia de Orta, Almada, Portugal*

**KEYWORDS:** adnexal mass; color Doppler; training; ultrasound

## ABSTRACT

***Objective*** *To estimate the interobserver agreement between a trainer and trainees in assigning the International Ovarian Tumor Analysis (IOTA) color score to adnexal masses using three-dimensional (3D) volumes and videoclips.*

***Methods*** *Fifty-one digital videoclips and 3D volumes of a non-consecutive series of adnexal masses were used for this study. One trainer and four trainees evaluated first the 3D volume and 1 week later a videoclip from the same mass. They had to assign IOTA color scores according to their impression of the amount of color content in each case. Interobserver agreement between trainer and trainees was assessed using Cohen's weighted kappa index with 95% CIs and percentage of agreement.*

***Results*** *When using 3D volumes, interobserver agreement was good for three out of four pairs of comparisons and very good for one (kappa values of 0.70, 0.68, 0.81 and 0.71 for trainees A, B, C and D, respectively). When using videoclips, interobserver agreement was very good for two out of four pairs of comparisons and good for two (kappa values of 0.84, 0.80, 0.68 and 0.86 for Trainees A, B, C and D, respectively).*

***Conclusion*** *Evaluation of IOTA color scores in adnexal masses using either videoclips or 3D volumes is reproducible even in the hands of trainees after a short training program. Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.*

## INTRODUCTION

Color/power Doppler ultrasound has been used extensively for assessing the vascularization of adnexal masses. Some reports have shown that adding color Doppler examination to gray-scale ultrasound may increase the specificity of the technique[1], while other reports have shown that it increases the confidence in diagnosis but not the diagnostic performance[2].

At present the use of color/power Doppler ultrasound is based on the assessment of flow location within a given mass, which is related to the features of the mass observed with gray-scale ultrasound[3], or it is based on the subjective quantification of the amount of flow[2]. Since this is a subjective assessment, the International Ovarian Tumor Analysis (IOTA) group suggested a classification system, the so-called 'color score system'[4]. According to this system, a color score of 1 means that no color signal is observed within the mass, a score of 2 means that a minimal amount of color signal is detected, a score of 3 means that a moderate amount of color signal is observed and a score of 4 means that abundant color Doppler signals are detected within the mass. This classification system is an important variable in the IOTA regression model 'LR1' for calculating the risk of malignancy of adnexal masses[5], and the finding of a color score of 1 is considered to indicate benignity (feature B5) while a color score of 4 is considered to indicate malignancy (feature M5) according to the IOTA simple rules[6].

To date, only two studies have assessed the interobserver agreement of IOTA color score, one of them using three-dimensional (3D) volumes and the other using digital videoclips[7,8]. Information regarding this issue is therefore still scanty.

We have developed a training program for the ultrasound assessment of adnexal masses mainly based on the use of stored 3D volumes[9]. We hypothesized that the evaluation of videoclips might be more reproducible than that of 3D volumes, and therefore potentially better for training, specifically for the assessment of color score. The objective of this study was to compare the interobserver agreement between a trainer and trainees

*Correspondence to*: Dr J. L. Alcázar, Department of Obstetrics and Gynecology, Clínica Universidad de Navarra, University of Navarra, Avenida Pio XII 36, 31008 Pamplona, Spain (e-mail: jlalcazar@unav.es)

*Accepted: 29 April 2014*

for assigning an IOTA color score to adnexal masses using 3D volumes and videoclips.

## METHODS

Fifty-one digital videoclips and 3D volumes of a non-consecutive series of adnexal masses were used for this study. Each mass had a videoclip and a 3D volume to be assessed. Cases were selected by the trainer (J.L.A.) from the training program database[9]. The selection was carried out based on 3D volume color content, with the aim of obtaining a similar number of cases with color scores of 1, 2, 3 and 4. The corresponding videoclip of the case was then retrieved from the database. Only cases with high-quality 3D volumes and videoclips available were selected. Institutional review board approval was obtained for the study.

Initially, the trainer reviewed all 3D volumes and then, a week later, their corresponding videoclips, evaluating the IOTA color score in each case in order to assess the intraobserver intermethod agreement for the trainer.

Prior to the evaluation of any volumes or videoclips, all trainees underwent a short (2 h) theoretical training session regarding the IOTA color score classification with exposure to several static images of cases with color score 1, 2, 3 or 4, as well as an introduction to all IOTA terms. Additionally, they all read the original paper in which the color score was described[4] and it was emphasized by the trainer that the recommendations suggested by Zannoni *et al.*[8] regarding how to estimate color content (only true tissue should be estimated and if there was uncertainty whether color signals belonged to the lesion or to the adjacent ovarian stroma they should be considered as belonging to the lesion) should be followed. All trainees were also instructed in the use of 4D View software (GE Healthcare Ultrasound, Milwaukee, WI, USA), specifically for virtual navigation and tomographic ultrasound imaging (TUI), since this software had to be used for manipulation of the 3D volumes. All trainees had had 3–6 months training on ultrasound in obstetrics and gynecology with no special focus on adnexal masses, but all were within our training program for ultrasound assessment of adnexal masses[9].

After the selection had been made, four trainees (Trainees A, B, C and D) evaluated the cases. First, 3D volumes were assessed by the trainees, blinded to each other's findings, using virtual navigation in multiplanar display and TUI. Rendering was not used. Each trainee evaluated all 3D volumes on different consecutive days (first Trainee A and last Trainee D) and was alone in the room in which evaluation was performed. They were instructed not to discuss their impressions among themselves or with the trainer after assessment. Each trainee had to assign an IOTA color score according to their impression of the amount of color content in each case.

**Table 1** Intraobserver intermethod agreement for the trainer in assigning color scores to 51 adnexal masses according to International Ovarian Tumor Analysis criteria using videoclips and three-dimensional (3D) volumes

| *Videoclip color score* | *3D volume color score* | | | |
|---|---|---|---|---|
| | *1* | *2* | *3* | *4* |
| 1 | 9 | — | — | — |
| 2 | 0 | 11 | 2 | 1 |
| 3 | 1 | 3 | 8 | 2 |
| 4 | — | — | — | 14 |

Data given as *n*.

One week later, each trainee assessed the videoclips to assign a color score in each case. The order in which videoclips were assessed was different from that of the 3D volumes. They had no access to the information regarding their assessment of the 3D volumes while evaluating the videoclips. Trainees were blinded to definitive diagnosis of the adnexal masses. We did not set a maximum time for performing evaluations of either 3D volumes or videoclips.

Additionally, after finishing the assessment the trainees had to answer the following questions:

1. In which method (3D volume or videoclip) do you feel you are most confident when assigning color score?
2. Which method (3D volume or videoclip) do you think is better for training in color score classification?

### Statistical analysis

Interobserver agreement between trainer and trainees was assessed using Cohen's weighted kappa index (κ) with

**Table 2** Ultrasound classification and definitive diagnosis for the 51 cases of adnexal masses included in the study

| *Ultrasound classification* | *Definitive diagnosis** |
|---|---|
| Unilocular (*n* = 20) | Dermoid cyst (*n* = 3)<br>Simple cyst (*n* = 5)<br>Endometrioma (*n* = 6)<br>Hemorrhagic cyst (*n* = 6) |
| Multilocular (*n* = 2) | Mucinous cystadenoma (*n* = 2) |
| Unilocular-solid (*n* = 8) | Cystadenofibroma (*n* = 2)<br>Primary ovarian cancer (*n* = 6) |
| Multilocular-solid (*n* = 10) | Metastatic cancer (*n* = 1)<br>Ovarian sarcoma (*n* = 1)<br>Mucinous cystadenoma (*n* = 2)<br>Borderline ovarian tumor (*n* = 2)<br>Primary ovarian cancer (*n* = 4) |
| Solid (*n* = 11) | Borderline ovarian tumor (*n* = 1)<br>Pelvic sarcoma (*n* = 1)<br>Metastatic cancer (*n* = 2)<br>Ovarian fibroma (*n* = 2)<br>Pedunculated myoma (*n* = 2)<br>Primary ovarian cancer (*n* = 3) |

*Diagnoses based on histopathological results after tumor removal, except for five of six hemorrhagic cysts and one of five simple cysts that resolved spontaneously after follow-up scan. Additionally, one dermoid cyst and two endometriomas were not surgically removed; diagnosis was based on ultrasound and the patients were being followed up at the time of writing.

**Table 3** Interobserver agreement between trainer and trainees using three-dimensional (3D) volumes and videoclips for the assessment of 51 adnexal masses according to International Ovarian Tumor Analysis criteria

| | *3D volume* | | *Videoclip* | |
|---|---|---|---|---|
| *Trainee* | *Weighted kappa (95% CI)* | *Agreement (%)* | *Weighted kappa (95% CI)* | *Agreement (%)* |
| Trainee A | 0.70 (0.55–0.85) | 69.8 | 0.84 (0.73–0.96) | 81.2 |
| Trainee B | 0.68 (0.54–0.85) | 65.8 | 0.80 (0.67–0.91) | 71.9 |
| Trainee C | 0.81 (0.69–0.92) | 76.7 | 0.68 (0.55–0.82) | 59.4 |
| Trainee D | 0.71 (0.55–0.86) | 69.8 | 0.86 (0.77–0.97) | 78.1 |

**Table 4** Interobserver agreement for International Ovarian Tumor Analysis color score between trainees using three-dimensional (3D) volumes and videoclips for the assessment of 51 adnexal masses

| | *Second trainee* | | | | | |
|---|---|---|---|---|---|---|
| | *3D volumes* | | | *Videoclips* | | |
| *First trainee* | *Trainee B* | *Trainee C* | *Trainee D* | *Trainee B* | *Trainee C* | *Trainee D* |
| Trainee A | 0.71<br>(0.57–0.86) | 0.60<br>(0.44–0.77) | 0.62<br>(0.45–0.78) | 0.79<br>(0.66–0.93) | 0.74<br>(0.60–0.88) | 0.59<br>(0.44–0.74) |
| Trainee B | — | 0.57<br>(0.40–0.73) | 0.60<br>(0.44–0.76) | — | 0.65<br>(0.48–0.83) | 0.51<br>(0.34–0.67) |
| Trainee C | — | — | 0.64<br>(0.48–0.80) | — | — | 0.65<br>(0.48–0.83) |

Data are given as weighted kappa (95% CI).

95% CIs and percentage of agreement[10]. A $\kappa$ of less than 0.20 indicates poor agreement; 0.21–0.40 indicates fair agreement; 0.41–0.60 indicates moderate agreement; 0.61–0.80 indicates good agreement; and 0.81–1.00 indicates very good agreement[11]. Considering that the expected percentage of agreement would be 70% and accepting a standard error of 20%, we calculated that a sample size of 50 cases would be needed. McNemar's test was used to compare overall color score assignments between videoclips and 3D volumes for the pooled data of all trainees[8].

## RESULTS

Intraobserver intermethod agreement for the trainer regarding IOTA color score as assessed using 3D volume and videoclip was very good ($\kappa = 0.82$ (95% CI, 0.74–0.95); percentage of agreement, 82.3%) (Table 1). The ultrasound diagnoses according to the IOTA classification system and the definitive diagnoses of the included cases are shown in Table 2.

Interobserver agreement between trainer and trainees using 3D volumes is shown in Table 3. Overall, it was good for three out of four pairs of comparisons and very good for one. Interobserver agreement between trainer and trainees for videoclip assessment is also shown in Table 3. Overall interobserver agreement was very good for two pairs of comparisons and good for two pairs. Interobserver agreement between trainees when using 3D volumes and videoclips is shown in Table 4, ranging from moderate to good.

When comparing all observations among trainees using 3D volumes or videoclips, we did not observe a statistically significant difference between the two methods (McNemar's test $P = 0.326$) (Table 5). All trainees reported that they felt that videoclips were better for training and also that they felt more confident in assigning color scores using videoclips.

## DISCUSSION

In this study we have shown that evaluation of IOTA color scores for adnexal masses is reproducible even in the hands of trainees after a short training program. The agreement between trainees and trainer is somewhat higher when using videoclips than when using 3D volumes for most pairs of comparisons.

The strength of our study is that all examiners evaluated the same cases using both 3D volumes and videoclips. Our results also suggest that if examiners follow the recommendations of Zannoni *et al.*[8] regarding how to assign a color score to adnexal masses it helps them in the evaluation and leads to better reproducibility.

**Table 5** Agreement between assignment of color scores in the assessment of 51 adnexal masses according to International Ovarian Tumor Analysis criteria using three-dimensional (3D) volumes and videoclips for the pooled results of all trainees

| | *3D volume color score* | | | |
|---|---|---|---|---|
| *Videoclip color score* | *1* | *2* | *3* | *4* |
| 1 | 25 | 8 | — | — |
| 2 | 9 | 27 | 14 | 6 |
| 3 | 1 | 15 | 25 | 18 |
| 4 | 1 | 3 | 8 | 44 |

Data are given as *n*.

Our data confirm previously reported results. Sladkevicius and Valentin reported that interobserver agreement for assigning color score was moderate (weighted $\kappa = 0.53$, percentage of agreement = 51%) using 3D volumes[7]. However, Zannoni *et al.*[8] reported that interobserver agreement was good (weighted κ ranged from 0.63 to 0.72 for all pairs of comparisons) when using videoclips. These two studies, like ours, have the limitation that neither of them evaluated interobserver agreement using real-time ultrasound. Therefore, the fact that the examiner is able to modify the color/power Doppler settings during real-time ultrasound was not taken into account, and we could not ascertain whether this factor could affect the subjective assessment of the amount of flow within the adnexal mass. Furthermore, interobserver reproducibility could be overestimated.

Our study has another significant limitation, as the time elapsed between the two evaluations was just 1 week and recall bias could exist. This could explain why most trainees had better agreement using videoclips than using 3D volumes.

We were rather surprised when observing that agreement for the trainees in our study was better than that observed in the studies of Zannoni *et al.* and Sladkevicius and Valentin, which were performed by trained examiners[7,8]. Several factors could potentially explain this. First, the short but intensive training program in our study immediately prior to assessment could have resulted in a better result than expected. Second, in our study cases with high-quality videoclips and 3D volumes available were selected. Thus a selection bias could also exist. Third, in our study the number of cases was smaller than those in the studies of Zannoni *et al.* and Sladkevicius and Valentin, so the number of 'discrepant' cases could be lower.

In spite of these limitations, we consider that our findings could have implications for training with respect to the ultrasound examination of adnexal masses. It is well known that experience substantially impacts on an examiner's diagnostic performance when evaluating adnexal masses by ultrasound[12], and that the subjective impression of an expert examiner is not improved upon by any other approach, such as the use of logistic regression models or scoring systems[13]. For this reason training in this field should be regarded as being of paramount importance. Some research has shown that theoretical training with static images does not improve diagnostic performance[14]. On the other hand, a long time might be needed for training based on real-time ultrasound. For this reason we developed a specific training program for the ultrasound evaluation of adnexal masses based on the off-line assessment of 3D volumes[9].

Trainees reported that they were more confident in assigning an IOTA score when using videoclips than when using 3D volumes. The use of videoclips could therefore be preferable for this task. However, it has to be borne in mind that optimal use of the 4D View software would require more training and supervision than that provided in this study. It is likely that further training in the use of this software would improve the confidence of the trainees.

## REFERENCES

1. Alcázar JL, Guerriero S, Laparte C, Ajossa S, Jurado M. Contribution of power Doppler blood flow mapping to gray-scale ultrasound for predicting malignancy of adnexal masses in symptomatic and asymptomatic women. *Eur J Obstet Gynecol Reprod Biol* 2011; **155**: 99–105.
2. Valentin L. Prospective cross-validation of Doppler ultrasound examination and gray-scale ultrasound imaging for discrimination of benign and malignant pelvic masses. *Ultrasound Obstet Gynecol* 1999; **14**: 273–283.
3. Guerriero S, Alcazar JL, Ajossa S, Galvan R, Laparte C, García-Manero M, Lopez-Garcia G, Melis GB. Transvaginal color Doppler imaging in the detection of ovarian cancer in a large study population. *Int J Gynecol Cancer* 2010; **20**: 781–786.
4. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I; International Ovarian Tumor Analysis (IOTA) Group. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound Obstet Gynecol* 2000; **16**: 500–505.
5. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L; International Ovarian Tumor Analysis Group. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *J Clin Oncol* 2005; **23**: 8794–8801.
6. Timmerman D, Testa AC, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, Paladini D, Van Calster B, Vergote I, Van Huffel S, Valentin L. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol* 2008; **31**: 681–690.
7. Sladkevicius P, Valentin L. Intra- and interobserver agreement when describing adnexal masses using the International Ovarian Tumor Analysis terms and definitions: a study on three-dimensional ultrasound volumes. *Ultrasound Obstet Gynecol* 2013; **41**: 318–327.
8. Zannoni L, Savelli L, Jokubkiene L, Di Legge A, Condous G, Testa AC, Sladkevicius P, Valentin L. Intra- and interobserver reproducibility of assessment of Doppler ultrasound findings in adnexal masses. *Ultrasound Obstet Gynecol* 2013; **42**: 93–101.
9. Alcázar JL, Díaz L, Flórez P, Guerriero S, Jurado M. Intensive training program for ultrasound diagnosis of adnexal masses: protocol and preliminary results. *Ultrasound Obstet Gynecol* 2013; **42**: 218–223.
10. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; **228**: 303–308.
11. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304**: 1491–1494.
12. Van Holsbeke C, Daemen A, Yazbek J, Holland TK, Bourne T, Mesens T, Lannoo L, Boes AS, Joos A, Van De Vijver A, Roggen N, de Moor B, de Jonge E, Testa AC, Valentin L, Jurkovic D, Timmerman D. Ultrasound experience substantially impacts on diagnostic performance and confidence when adnexal masses are classified using pattern recognition. *Gynecol Obstet Invest* 2010; **69**: 160–168.
13. Timmerman D. The use of mathematical models to evaluate pelvic masses; can they beat an expert operator? *Best Pract Res Clin Obstet Gynaecol* 2004; **18**: 91–104.
14. Van Holsbeke C, Daemen A, Yazbek J, Holland TK, Bourne T, Mesens T, Lannoo L, De Moor B, De Jonge E, Testa AC, Valentin L, Jurkovic D, Timmerman D. Ultrasound methods to distinguish between malignant and benign adnexal masses in the hands of examiners with different levels of experience. *Ultrasound Obstet Gynecol* 2009; **34**: 454–461.