*Ultrasound Obstet Gynecol* 2014; **44**: 95–99
Published online 21 May 2014 in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.13254

# Interobserver agreement in describing adnexal masses using the International Ovarian Tumor Analysis simple rules in a real-time setting and using three-dimensional ultrasound volumes and digital clips

B. RUIZ DE GAUNA*, P. SANCHEZ†, L. PINEDA‡, J. UTRILLA-LAYNA‡, L. JUEZ‡ and J. L. ALCÁZAR‡

**Barcelona Center for Maternal Fetal and Neonatal Medicine, Hospital Sant Joan de Deu and Hospital Clinic, Universitat de Barcelona, Barcelona, Spain; †Department of Obstetrics and Gynecology, Hospital Fundación Jimenez Diaz, Madrid, Spain; ‡Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, Pamplona, Spain*

KEYWORDS: adnexal mass; reproducibility; simple rules; ultrasound

## ABSTRACT

***Objective*** *To estimate the agreement between an expert and a non-expert examiner using the International Ovarian Tumor Analysis (IOTA) simple rules for classifying adnexal masses on real-time ultrasound and when using three-dimensional (3D) ultrasound volumes and digital clips.*

***Methods*** *Forty-two non-consecutive women diagnosed as having an adnexal mass were evaluated by transvaginal power Doppler ultrasound as part of their diagnostic work-up. In each woman, examination was first performed by a non-expert examiner (a trainee) and immediately afterwards by an expert examiner. Both used the IOTA simple rules to describe the mass, blinded to each other's results. After finishing the examination, each examiner classified the mass as benign, malignant or inconclusive, according to the IOTA simple rules. Additionally, the expert recorded a short videoclip and acquired a static 3D volume of each mass, which were subsequently assessed by four trainees in obstetrics and gynecology with different levels of training, who also classified the mass as benign, malignant or inconclusive according to the IOTA simple rules. Agreement was assessed by calculating weighted and standard kappa index values with 95% CI and the percentage of agreement between observers.*

***Results*** *Agreement between the observers who performed real-time ultrasound examination was good (weighted kappa = 0.76; 95% CI, 0.61–0.90; agreement = 78.6%). Agreement between trainees using videoclips plus 3D volumes was moderate (kappa values ranged from 0.45 to 0.58, depending on pair comparison).*

***Conclusion*** *Interobserver agreement of the IOTA simple rules for classifying adnexal masses as benign, malignant or inconclusive using real-time ultrasound, between an expert and a non-expert examiner, might be considered good. Agreement using a videoclip plus a 3D volume was moderate for trainees with different degrees of training. Copyright © 2013 ISUOG. Published by John Wiley & Sons Ltd.*

## INTRODUCTION

In 2008, the International Ovarian Tumor Analysis (IOTA) group proposed the so-called 'simple rules' for ultrasound classification of adnexal masses[1]. The use of this approach is appealing because the simple rules are based on the identification of basic features of the adnexal mass during ultrasound examination. It has been reported that the simple rules can be applied in about 75–80% of all adnexal masses[2–4]. When the mass can be classified as benign or malignant using the simple rules, the diagnostic performance is good, even when applied by examiners with differing levels of expertise[5,6].

However, to the best of our knowledge, no study has assessed the agreement between observers for classifying adnexal masses using the IOTA simple rules on real-time ultrasound. In diagnostic imaging, estimating the agreement between different observers is crucial for a given diagnostic method to be introduced into clinical practice.

The aims of this study were threefold: first, to estimate agreement on real-time ultrasound examination between one experienced and one inexperienced ultrasound examiner with regard to classifying adnexal masses as benign, malignant or unclassifiable using the IOTA simple rules;

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

*Accepted: 15 November 2013*

second, to estimate agreement on real-time ultrasound examination between one experienced and one inexperienced ultrasound examiner with regard to the presence of each of the 10 ultrasound features included in the IOTA simple rules; and, third, to estimate agreement between trainees with different levels of ultrasound experience with regard to classifying adnexal masses as benign, malignant or unclassifiable using the IOTA simple rules when evaluating three-dimensional (3D) ultrasound volumes and digital clips.

## METHODS

This was a prospective observational study, performed in a tertiary care university hospital. Non-consecutive women diagnosed as having a persistent adnexal mass evaluated during a 2-month period (December 2012 to January 2013) were included in the study. Institutional Review Board approval was obtained and all women gave oral informed consent. Patient selection was performed according to two criteria: first, the availability of the expert examiner and trainee at the ultrasound unit; and, second, the whole mass or most of the mass could be included in a single 3D volume.

Observers included an expert examiner (J.L.A., with more than 20 years' experience in gynecologic ultrasound), a 3rd-year trainee in obstetrics and gynecology for real-time ultrasound (B.R.G., this trainee had a formal 3-month period of real-time ultrasound training, under the supervision of an expert examiner) and four trainees in obstetrics and gynecology for assessment of videoclips and 3D volumes (P.S., a 4th-year resident with 6 months' formal training in gynecological ultrasound; L.P., a 3rd-year resident with 3 months' formal training in gynecological ultrasound; J.U.L., a 2nd-year resident with 1 month of formal training in gynecological ultrasound; and L.J., a 1st-year resident with no training in practical ultrasound but who had undergone a theoretical course, including the IOTA simple rules).

All trainees learnt about the IOTA simple rules by reading the original paper published by the IOTA group[1]. Additionally, the expert examiner gave a lecture to all trainees about the IOTA simple rules, showing representative images of each ultrasound feature, before the start of the study.

All women were evaluated by transvaginal ultrasound as part of our routine diagnostic work-up using a Voluson E8 equipped with a 5–9-MHz endovaginal probe, power Doppler and 3D/four-dimensional (4D) ultrasound facilities (GE Healthcare Ultrasound, Milwaukee, WI, USA). Ultrasound examination was first performed by the 3rd-year resident in obstetrics and gynecology (B.R.G.) and, immediately afterwards, the expert examiner (J.L.A.) also carried out an ultrasound examination. Examiners were blinded to each other's results. Both examiners had to evaluate the presence or absence of each benign or malignant ultrasound feature and to classify the mass as benign, malignant or inconclusive according to the IOTA simple rules[1] (Table 1), recording the findings in an Excel datasheet (Microsoft Inc., New York, NY, USA) for subsequent analysis.

After the expert finished the real-time ultrasound evaluation he recorded a short videoclip (about 15–20 s) and acquired a static 3D volume of the mass. The video and 3D volume from each mass were subsequently assessed by the four trainees in obstetrics and gynecology. These trainees had to classify the mass as benign, malignant or inconclusive according to the IOTA simple rules, looking first at the videoclip and then manipulating the 3D volume using the 4DView$^{TM}$ software (GE Healthcare Ultrasound). All four of these examiners were blinded to each other's results and to the results of the real-time ultrasound examinations.

### Statistical analysis

Agreement was estimated by calculating the weighted kappa index[7] and the percentage of agreement in classifying the mass as benign, inconclusive or malignant.

We also assessed the agreement for each ultrasound feature between examiners performing real-time ultrasound by calculating the standard kappa index with 95% CI[8]. A kappa value of < 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[9]. GraphPad QuickCalcs software was used to calculate the kappa and weighted indices (GraphPad Software Inc., La Jolla, CA, USA). Power and sample size estimations were not performed.

## RESULTS

Forty-two women, mean age 35.6 (SD, 11.6; range, 21–68) years, were included in the study. In 34 of 42 cases, the mass was removed surgically and histological diagnosis was available (borderline ovarian tumors, $n = 4$; primary ovarian cancer, $n = 9$; metastatic ovarian cancer, $n = 1$; endometrioma, $n = 8$; ovarian fibroma, $n = 3$; serous cystadenoma, $n = 2$; dermoid cyst, $n = 2$; serous cystadenofibroma, $n = 2$; mucinous cystadenofibroma, $n = 1$; para-ovarian cyst, $n = 1$; and granulosa cell tumor, $n = 1$). In the eight remaining cases (in which the mass appeared benign), women were followed up with further examinations (presumed diagnoses: four hemorrhagic cysts, three endometriomas and one simple cyst).

Agreement for classifying the mass as benign, malignant or inconclusive, based on the IOTA simple rules between expert examiner and non-expert examiner on real-time ultrasound assessment, was good (weighted kappa = 0.76; 95% CI, 0.61–0.90; percentage of agreement = 78.6%) (Table 2). Agreement for each benign or malignant ultrasound feature during real-time ultrasound is shown in Table 3. Agreement was very good for the features 'unilocular tumor', 'smooth multilocular tumor with largest diameter < 100 mm' and 'irregular multilocular solid tumor with largest diameter ≥ 100 mm'; good for 'presence of solid components where solid component largest diameter is < 7 mm', 'at least four papillary

**Table 1** Simple rules for classifying adnexal masses proposed by the International Ovarian Tumor Analysis (IOTA) group[1]

| *Features for malignant tumor* | | *Features for benign tumor* | |
|---|---|---|---|
| M1 | Irregular solid tumor | B1 | Unilocular tumor |
| M2 | Presence of ascites | B2 | Presence of solid components where solid component's largest diameter < 7 mm |
| M3 | At least four papillary projections | B3 | Presence of acoustic shadows |
| M4 | Irregular multilocular solid tumor with largest diameter ≥ 100 mm | B4 | Smooth multilocular tumor with largest diameter < 100 mm |
| M5 | Very strong blood flow (color score 4) | B5 | No blood flow (color score 1) |

**Table 2** Agreement between expert examiner and trainee on real-time ultrasound examination with regard to classifying adnexal masses as benign, malignant or inconclusive using the International Ovarian Tumor Analysis (IOTA) simple rules

| | *Trainee* | | | |
|---|---|---|---|---|
| | *Benign* | *Inconclusive* | *Malignant* | *Total* |
| Expert examiner | | | | |
| Benign | 17 | 0 | 0 | 17 |
| Inconclusive | 4 | 2 | 2 | 8 |
| Malignant | 1 | 2 | 14 | 17 |
| Total | 22 | 4 | 16 | 42 |

Data are given as *n*. Weighted kappa[8] = 0.76 (95% CI, 0.61–0.90); percentage agreement = 78.6% (33 of 42).

projections' and 'very strong blood flow (color score 4)'; moderate for 'no blood flow (color score 1)', 'presence of ascites' and 'irregular solid tumor'; and fair for 'presence of acoustic shadows'.

Agreement between trainees classifying the mass as benign, inconclusive or malignant when assessing videoclips plus 3D volumes was moderate (Table 4).

## DISCUSSION

In this study we found that agreement with regard to classifying adnexal masses as benign, malignant or inconclusive using the IOTA ultrasound-based simple rules between an expert and a less-experienced examiner on real-time ultrasound is good. When we focused on each ultrasound feature we observed that agreement was very good, good or moderate for most of them, but that agreement beyond that expected by chance was only fair for the presence of acoustic shadows.

The main strength of our study is that, to the best of our knowledge, this is the first to assess interobserver agreement with regard to describing adnexal masses using the IOTA simple rules during real-time ultrasound. We also estimated the agreement among trainees, with different levels of training in ultrasound, for classifying adnexal masses as benign, malignant or inconclusive using the IOTA simple rules applied to videoclips and stored 3D volumes. We found that agreement amongst them was moderate.

A limitation of this study is that the series is small and comprises a selected population. We did not perform sample size estimation, and the 95% CIs for the kappa index are wide, so estimation may be imprecise. Additionally, the evaluation of interobserver agreement among non-expert examiners using videoclips plus 3D ultrasound should be considered suboptimal, as all examined the same stored data and, as a result, the analysis does not include inherent sources of variability between different acquisitions. It is therefore possible that the estimated interobserver agreement was overestimated.

We consider our findings are clinically relevant because recent studies have shown that IOTA ultrasound-based simple rules perform well in the hands of examiners with different degrees of experience and training[10]. However, the lack of consistency between observers has long been recognized as a problem in clinical diagnosis[9]. If a diagnostic approach using imaging has a good performance, but is not reproducible among observers, then its use in clinical practice could be questioned. Therefore, assessing the reliability and consistency of the method is essential[7]. Our study confirms reproducibility of results among observers with different levels of experience when using the IOTA simple rules for classifying adnexal masses.

We observed that the identification of two relatively simple features, 'the presence of acoustic shadows' and 'the presence of ascites', showed the worst agreement between expert and trainee on real-time imaging. One could argue that both features had a low prevalence and that this could affect the results because disagreement in just a couple of cases could lead to a low kappa index. However, this was also the case for other features, such as 'presence of solid component < 7 mm' or 'four or more papillary projections', in which the kappa index was higher.

Acoustic shadowing is dependent on the examiner's impression, and this could explain the discrepancy between observers. Sladkevicius and Valentin found that agreement for the presence of acoustic shadows was good[11]. In this study, both examiners were expert examiners. In our study, one examiner was non-expert and the other was expert. It could be speculated that different expertise could affect the interpretation of acoustic shadowing, making agreement worse.

Regarding identification of 'the presence of ascites', we were surprised by the low interobserver agreement found. We used the IOTA definition of ascites as fluid outside the pouch of Douglas. In our opinion, this is also a rather subjective definition and expertise may also affect agreement. In fact, we noticed that the non-expert examiner overestimated the presence of ascites.

**Table 3** Agreement between expert examiner and trainee on realtime ultrasound for each ultrasound feature included in the International Ovarian Tumor Analysis (IOTA) simple rules

| *Feature* | *Prevalence %* *(n/n)** | *Kappa index* *(95% CI)* | *Percentage agreement* *(n/n)* |
|---|---|---|---|
| B1: unilocular tumor | 17.8 (15/84) | 0.89 (0.75 to 1.0) | 95.2 (40/42) |
| B2: presence of solid components where solid component's largest diameter < 7 mm | 4.8 (4/84) | 0.64 (0.19 to 1.0) | 95.2 (40/42) |
| B3: presence of acoustic shadows | 4.8 (4/84) | 0.36 (−0.20 to 0.92) | 92.9 (39/42) |
| B4: smooth multilocular tumor with largest diameter < 100 mm | 8.3 (7/84) | 0.91 (0.73 to 1.0) | 97.6 (41/42) |
| B5: no blood flow (color score 1) | 19.0 (16/84) | 0.53 (0.25 to 0.82) | 80.9 (34/42) |
| M1: irregular solid tumor | 14.3 (12/84) | 0.58 (0.27 to 0.88) | 85.7 (36/42) |
| M2: presence of ascites | 10.7 (9/84) | 0.42 (0.03 to 0.80) | 85.7 (36/42) |
| M3: at least four papillary projections | 4.8 (4/84) | 0.64 (0.19 to 1.0) | 95.2 (40/42) |
| M4: irregular multilocular solid tumor with largest diameter ≥ 100 mm | 8.3 (7/84) | 0.81 (0.55 to 1.0) | 95.2 (40/42) |
| M5: very strong blood flow (color score 4) | 28.6 (24/84) | 0.63 (0.40 to 0.85) | 80.9 (34/42) |

*Number of times the feature was observed by any observer/total number of observations.

**Table 4** Agreement between trainees with regard to classifying adnexal masses as benign, malignant or unclassifiable on the basis of the International Ovarian Tumor Analysis (IOTA) simple rules when analyzing videoclips and three-dimensional volumes

| | *4th-year trainee* | *3rd-year trainee* | *2nd-year trainee* | *1st-year trainee* |
|---|---|---|---|---|
| 4th-year trainee | — | WK = 0.56 (0.33–0.80)<br>[76.2%] | WK = 0.58 (0.35–0.81)<br>[76.2%] | WK = 0.51 (0.27–0.73)<br>[71.4%] |
| 3rd-year trainee | — | — | WK = 0.51 (0.26–0.74)<br>[72.2%] | WK = 0.46 (0.21–0.71)<br>[71.4%] |
| 2nd-year trainee | — | — | — | WK = 0.45 (0.20–0.71)<br>[71.4%] |
| 1st-year trainee | — | — | — | — |

Data are given as weighted kappa (WK) (95% CI) [percentage agreement].

It could seem surprising that agreement for two rather complicated features, 'irregular multilocular solid tumor with largest diameter ≥ 100 mm' and 'smooth multilocular tumor with largest diameter < 100 mm', was good. We do not have a clear explanation for this finding. It is probable that these features, considered to be more difficult to assess by the expert examiner, received more emphasis during training, resulting in a higher agreement.

Sladkevicius and Valentin also assessed interobserver agreement for the grayscale variable 'irregular/smooth surface' (involved in features B4 and M4). They found that agreement was moderate[11].

According to our results, agreement for color score (features B5 and M5) was good. This could be considered to be in agreement with the results reported by Zannoni and coworkers. In this study, seven examiners with different levels of experience assessed the reproducibility of the IOTA color score in a series of 103 digital videoclips from adnexal masses[12]. They found that interobserver agreement for the four different results of the IOTA color score was good. However, Sladkevicius and Valentin, using stored 3D volumes, found that interobserver agreement was just moderate[11]. However, it should be borne in mind that the chances of disagreement are higher when four options, instead of just two results, are possible, as used in our study. Additionally, assessing the amount of color is also subjective and dependent on the ability to adjust correctly Doppler settings and understand the pitfalls of color Doppler imaging. In our study both examiners used the same color Doppler settings.

As stated above, we found moderate agreement for classifying adnexal masses, using the IOTA simple rules, among trainees with different levels of training. Guerriero *et al.* recently published a study estimating inter- and intraobserver agreement in the classification of adnexal masses by applying the IOTA simple rules to stored 3D volumes[13]. They used 100 stored 3D volumes that were analyzed by five different examiners (two expert examiners, one moderate expert and two trainees). Consistent with our results, they found that interobserver agreement was moderate. In our opinion, all of these findings highlight the relevance of adequate training for trainees.

We consider that our results could be generalizable, at least for examiners with similar levels of ultrasound experience and who have received education on the IOTA simple rules similar to that in this study. Additionally, our results could be generalizable for a population with similar tumor characteristics. It is likely that a study population including many tumors with equivocal ultrasound features would yield different results.

Although our series is small, the histological distribution of masses (benign/malignant) is similar to that of a larger IOTA study assessing ultrasound-based simple rules[10]. However, we must be cautious with our conclusions because our estimates may be imprecise, both for real-time ultrasound examination and for the use of videoclips and 3D volumes.

## REFERENCES

1. 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.
2. Timmerman D, Ameye L, Fischerova D, Epstein E, Melis GB, Guerriero S, Van Holsbeke C, Savelli L, Fruscio R, Lissoni AA, Testa AC, Veldman J, Vergote I, Van Huffel S, Bourne T, Valentin L. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. *BMJ* 2010; **341:** c6839.
3. Fathallah K, Huchon C, Bats AS, Metzger U, Lefrère-Belda MA, Bensaid C, Lécuru F. External validation of simple ultrasound rules of Timmerman on 122 ovarian tumors. *Gynecol Obstet Fertil* 2011; **39:** 477–481.
4. Hartman CA, Juliato CR, Sarian LO, Toledo MC, Jales RM, Morais SS, Pitta DD, Marussi EF, Derchain S. Ultrasound criteria and CA 125 as predictive variables of ovarian cancer in women with adnexal tumors. *Ultrasound Obstet Gynecol* 2012; **40:** 360–366.
5. Sayasneh A, Wynants L, Preisler J, Kaijser J, Johnson S, Stalder C, Husicka R, Abdallah Y, Raslan F, Drought A, Smith AA, Ghaem-Maghami S, Epstein E, Van Calster B, Timmerman D, Bourne T. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer* 2013; **108:** 2448–2454.
6. Alcázar JL, Pascual MA, Olartecoechea B, Graupera B, Aubá M, Ajossa S, Hereter L, Julve R, Gastón B, Peddes C, Sedda F, Piras A, Saba L, Guerriero S. IOTA simple rules for discriminating between benign and malignant adnexal masses: a prospective external validation. *Ultrasound Obstet Gynecol* 2013; **42:** 467–471.
7. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; **228:** 303–308.
8. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20:** 37–46.
9. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304:** 1491–1494.
10. Sayasneh A, Kaijser J, Preisler J, Johnson S, Stalder C, Husicka R, Guha S, Naji O, Abdallah Y, Raslan F, Drought A, Smith AA, Fotopoulou C, Ghaem-Maghami S, Van Calster B, Timmerman D, Bourne T. A multicenter prospective external validation of the diagnostic performance of IOTA simple descriptors and rules to characterize ovarian masses. *Gynecol Oncol* 2013; **130:** 140–146.
11. 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.
12. 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.
13. Guerriero S, Saba L, Ajossa S, Peddes C, Sedda F, Piras A, Olartecoechea B, Aubá M, Alcázar JL. Assessing the reproducibility of the IOTA simple ultrasound rules for classifying adnexal masses as benign or malignant using stored 3D volumes. *Eur J Obstet Gynecol Reprod Biol* 2013; **171:** 157–160.

This article has been selected for Journal Club.

A slide presentation, prepared by Dr Tommaso Bignardi, one of UOG's Editors for Trainees, is available online.