Ultrasound Obstet Gynecol 2014; **44**: 100–108
Published online 1 June 2014 in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.13273

Intra- and interobserver agreement with regard to describing adnexal masses using International Ovarian Tumor Analysis terminology: reproducibility study involving seven observers

L. ZANNONI*†, L. SAVELLI*, L. JOKUBKIENE†, A. DI LEGGE‡, G. CONDOUS§, A. C. TESTA‡, P. SLADKEVICIUS† and L. VALENTIN†

**Department of Obstetrics and Gynecology, S. Orsola Malpighi Hospital, University of Bologna, Bologna, Italy; †Department of Obstetrics and Gynecology, Skåne University Hospital Malmö, Lund University, Malmö, Sweden; ‡Department of Obstetrics and Gynecology, Catholic University of the Sacred Heart, Rome, Italy; §Acute Gynaecology, Early Pregnancy and Advanced Endosurgery Unit, Sydney Medical School Nepean, University of Sydney, Sydney, Australia*

KEYWORDS: IOTA; ovarian neoplasms; reproducibility of results; ultrasonography

ABSTRACT

Objectives *To estimate intraobserver repeatability and interobserver agreement in assessing the presence of papillary projections in adnexal masses and in classifying adnexal masses using the International Ovarian Tumor Analysis terminology for ultrasound examiners with different levels of experience. We also aimed to identify ultrasound findings that cause confusion and might be interpreted differently by different observers, and to determine if repeatability and agreement change after consensus has been reached on how to interpret 'problematic' ultrasound images.*

Methods *Digital clips (two to eight clips per adnexal mass) with gray-scale and color/power Doppler information of 83 adnexal masses in 80 patients were evaluated independently four times, twice before and twice after a consensus meeting, by four experienced and three less experienced ultrasound observers. The variables analyzed were tumor type (unilocular, unilocular solid, multilocular, multilocular solid, solid) and presence of papillary projections. Intraobserver repeatability was evaluated for each observer (percentage agreement, Cohen's kappa). Interobserver agreement was estimated for all seven observers (percentage agreement, Fleiss kappa, Cohen's kappa).*

Results *There was uncertainty about how to define a solid component and a papillary projection, but consensus was reached at the consensus meeting. Interobserver agreement for tumor type was good both before and after the consensus meeting, with no clear improvement after the meeting, mean percentage agreement being 76.0% (Fleiss kappa, 0.695) before the meeting and 75.4% (Fleiss kappa, 0.682) after the meeting. Interobserver agreement with regard to papillary projections was moderate both before and after the consensus meeting, with no clear improvement after the meeting, mean percentage agreement being 86.6% (Fleiss kappa, 0.536) before the meeting and 82.7% (Fleiss kappa, 0.487) after it. There was substantial variability in pairwise agreement for papillary projections (Cohen's kappa, 0.148–0.787). Intraobserver repeatability with regard to tumor type was very good and similar before and after the consensus meeting (agreement 87–95%, kappa, 0.83–0.94). With regard to papillary projections intraobserver repeatability was good or very good both before and after the consensus meeting (agreement 88–100%, kappa, 0.64–1.0).*

Conclusions *Despite uncertainty about how to define solid components, interobserver agreement was good for tumor type. The interobserver agreement for papillary projection was moderate but very variable between observer pairs. The term 'papillary projection' might need a more precise definition. The consensus meeting did not change inter- or intraobserver agreement.* *Copyright © 2013 ISUOG. Published by John Wiley & Sons Ltd.*

INTRODUCTION

Conservative management or minimally invasive conservative surgery should be offered to patients with ovarian or other adnexal masses only if the mass can be confidently classified as benign[1–3]. Subjective assessment of an adnexal mass by an experienced ultrasound examiner has been shown to be the best method for discriminating

Correspondence to: Dr L. Zannoni, Department of Obstetrics and Gynecology, S.Orsola Malpighi Hospital, via Massarenti 13, 40138 Bologna, Italy (e-mail: letizia.zannoni@studio.unibo.it)

Accepted: 22 November 2013

between benign and malignant adnexal masses[4,5]. Because it is impossible for every patient to be examined by an expert sonologist, mathematical models to estimate the risk of malignancy have been developed with the hope of replicating the performance of an experienced ultrasound operator[6]. The International Ovarian Tumor Analysis (IOTA) logistic regression models LR1 and LR2 have been shown to be superior to all other models for this purpose and have been recommended for use in clinical practice[6–8]. LR1 and LR2 include six and 12 variables, respectively, most of which are ultrasound variables[7]. Definitions of these ultrasound variables have been published[9]. However, even when using standardized terms and definitions, ultrasound examiners may evaluate the features of an adnexal mass differently, which means that the risk of malignancy calculated by LR1 or LR2 may vary both within and between examiners[10]. One of the variables in LR1 and LR2 is the color score, a score based on subjective estimation of the color content of the tumor scan. In a reproducibility study in which seven observers with different levels of ultrasound experience estimated the color content of tumor scans from digital videoclips, it became clear that different observers interpret the images differently. Factors causing uncertainty about how to estimate the color content were identified, and interobserver agreement improved slightly after consensus had been reached among the observers on how to interpret the color Doppler images[11]. Intraobserver repeatability and interobserver agreement when using the IOTA terminology to describe gray-scale ultrasound images of adnexal masses have been published for one experienced observer pair[10]. However, to the best of our knowledge, the problems that ultrasound examiners may encounter when they use the IOTA terminology to describe gray-scale ultrasound images of adnexal masses have not been elucidated in any published study.

The aims of this study were to estimate intraobserver repeatability and interobserver agreement in assessing the presence of papillary projections in adnexal masses and in classifying adnexal masses using the IOTA terminology for ultrasound examiners with different levels of experience. Also we aimed to identify ultrasound findings that cause confusion and might be interpreted differently by different observers, and to determine whether intraobserver repeatability and interobserver agreement change after consensus has been reached on how to interpret 'problematic' ultrasound images.

METHODS

This was a prospective observational study set in four university hospitals. Seven ultrasound examiners (observers) participated in the study, four being gynecologists from tertiary referral gynecological ultrasound units who had more than 10 years' experience in gynecological ultrasonography and a special interest in adnexal masses (Observers A (L.V.), B (L.S.), C (P.S.) and D (A.C.T)). Observer E (G.C.) was a senior gynecologist, very skilled in gynecological ultrasound but with a field of interest other than adnexal masses and there were two trainees (Observers F (L.J.) and G (A.D.L.)), who had received at least 2 years' training in gynecological ultrasound in the ultrasound departments of Observers A and D, respectively.

The adnexal masses were classified as unilocular, unilocular solid, multilocular, multilocular solid or solid using IOTA terminology, and the presence or absence of papillary projections in adnexal masses was also defined according to IOTA terminology[9].

Gray-scale and color/power Doppler digital clips of adnexal masses of 80 of the 100 patients included in a study on the reproducibility of the IOTA color score, i.e. those 80 patients with available digital gray-scale ultrasound clips of their adnexal masses, comprised the source of our data[11].

The clips were collected by three gynecologists skilled in ultrasonography (Observers A, B and C; see above). High-end ultrasound systems with high-frequency transducers were used in all cases. After completion of the scanning phase of the study, all clips were checked by the first author, who did not participate in either the collection or the evaluation of the clips. Acceptable clips (i.e. those in which morphological and color/power Doppler features of the whole mass were seen and the duration of the clip was at least 4 s) were copied to DVDs. For most adnexal masses more than one 4-s clip was collected and sometimes clips from the same tumor were collected with two different ultrasound machines. The DVDs were distributed to the seven ultrasound examiners selected to analyze the clips. The time between collection of the clips and first analysis of them varied between 1 month and 5 years.

The observers were instructed to read the article describing the IOTA terms and definitions to be used when describing ultrasound images of adnexal masses before starting to evaluate the clips[9]. Each observer analyzed the clips four times: twice before a planned consensus meeting (analyses 1 and 2) and twice after they had attended the consensus meeting (analyses 3 and 4). The DVDs had been prepared so that the order in which the video clips were presented was different for analyses 1 and 2 and different for analyses 3 and 4. The time between analyses 1 and 2 was at least 2 weeks, as was that between analyses 3 and 4 and that between analyses 2 and 3. Each observer first assessed the gray-scale clips of each patient. Using the IOTA terms and definitions each mass was classified as unilocular, unilocular solid, multilocular, multilocular solid or solid and the presence or absence of papillary projections was noted[9]. The results of the evaluation of the gray-scale clips of each patient were immediately recorded in a dedicated research form before proceeding to assess the color Doppler clips of the same patient. The observers were blinded to each other's results and to their own previous results, and all observers were unaware of the postoperative histopathological diagnosis of the masses.

After all seven examiners had completed analyses 1 and 2 of the clips they participated in a 1-day consensus meeting to discuss the problems that they had

encountered when analyzing the clips. The consensus meeting was divided into two parts. In the first part the observers evaluated independently of each other the gray-scale and color/power Doppler digital clips of 30 adnexal masses collected by the first author, but not included in the DVDs (thus none of the examiners had seen these clips before). Each observer noted their evaluation in a dedicated research form. In the second part, the discrepancies between the observers when evaluating these 30 clips were identified from the forms and discussed, after which the examiners reached consensus on how to classify adnexal masses with regard to tumor type and presence of papillations. The seven observers were asked to stick to the consensus agreement during analyses 3 and 4 of the clips to be performed after the consensus meeting.

Statistical analysis

Interobserver agreement was estimated by calculating the percentage agreement for each observer pair and then averaging the results (mean percentage agreement), by calculating Fleiss kappa for multiple observers (jack-knife estimates) and Cohen's kappa for each observer pair, with mean and range of Cohen's kappa values reported[12,13]. Results are presented for all seven observers as well as for the four most experienced observers and the three less experienced observers separately. Intraobserver repeatability was estimated for each of the seven observers, results being expressed as percentage agreement and Cohen's kappa[13]. Interobserver agreement and intraobserver repeatability before the consensus meeting were compared with those after the consensus meeting.

Kappa values are a measure of by how much the observed agreement exceeds agreement by chance, and can be calculated only if field tables are symmetrical. Kappa values tend to be low if data are very skewed, even if agreement is close to 100% (e.g. if two observers agree that a particular finding is absent in 90% of cases and present in 5% of cases)[14]. It has been suggested that kappa values of 0.81–1.0 may be taken to indicate very good agreement, 0.61–0.80 good agreement, 0.41–0.60 moderate agreement, 0.21–0.40 fair agreement and values ≤ 0.20 poor agreement[15].

Statistical analysis for intraobserver agreement was carried out using the SPSS version 16 (SPSS Inc., Chicago, IL, USA) and those for interobserver agreement using a small SPSS-program (macro) from http://www.ccitonline.org/jking/homepage/ and SPSS version 21.

RESULTS

Our study included videoclips with both gray-scale and color Doppler information from 83 adnexal masses in 80 women. For eight (10%) masses, clips from the same tumor had been collected with two different ultrasound machines. The mean age of the 80 patients (three having bilateral masses) contributing adnexal masses to our study was 47 (range, 19–92) years, four (5%) patients were pregnant, and 31 (39%) were postmenopausal. Table 1 shows the histological diagnoses of the 83 masses included.

Table 1 Histological diagnoses of 83 adnexal masses included in the study

Diagnosis	n
Benign masses	
Ovarian	
Endometrioma	9
Teratoma	11
Serous cystadenoma	4
Mucinous cystadenoma	3
Cystadenofibroma	5
Fibrothecoma	2
Ovarian fibroma	5
Struma ovarii	1
Luteal cyst	1
Simple cyst	3
Non-ovarian	
Simple paraovarian cyst	3
Paraovarian cystadenoma	1
Hydrosalpinx	1
Pedunculated myoma	2
Borderline masses	
Ovarian	3
Paraovarian	1
Primary invasive masses	
Ovarian	
Papillary serous cystadenocarcinoma	7
Endometrioid adenocarcinoma	5
Clear-cell carcinoma	1
Granulosa-cell tumor	2
Sertoli Leydig cell tumor	1
Dysgerminoma	1
Immature cystic teratoma	1
Type not specified	4
Non-ovarian	
Tubal serous cystadenocarcinoma	1
Intestinal leiomyosarcoma	1
Metastases in the ovary from:	
Breast cancer	2
Unknown primary tumor	2

Interobserver agreement with regard to tumor type is shown in Table 2. Agreement (beyond chance between all seven observers) was good both before and after the consensus meeting, with no clear improvement after the meeting. The average percentage agreement was 76.0% (Fleiss kappa, 0.695) before the consensus meeting and 75.4% (Fleiss kappa, 0.682) after the meeting. Before the meeting, interobserver agreement beyond chance between the four experienced observers was better than that between the three less experienced observers. After the meeting the agreement beyond chance between the three less experienced observers improved substantially, while no improvement was noted for the four most experienced observers.

Interobserver agreement with regard to the presence of papillary projections is shown in Table 3. The agreement beyond chance between all seven observers was moderate

Table 2 Interobserver agreement in 83 cases of adnexal mass with regard to tumor type (unilocular, unilocular solid, multilocular, multilocular solid, solid) among seven observers before and after a consensus meeting

Parameter	*Agreement (%)*	*Fleiss kappa*	*Cohen's kappa*
First evaluation before meeting			
All observers ($n = 7$)	76.0 (60.2–85.5)	0.695 (0.687–0.703)	0.690 (0.500–0.810)
Experienced observers ($n = 4$)	78.7 (72.3–81.9)	0.735 (0.726–0.744)	0.727 (0.653–0.762)
Less experienced observers ($n = 3$)	70.6 (67.5–74.7)	0.621 (0.611–0.632)	0.617 (0.576–0.672)
First evaluation after meeting			
All observers ($n = 7$)	75.4 (66.3–85.5)	0.682 (0.675–0.689)	0.682 (0.554–0.812)
Experienced observers ($n = 4$)	73.5 (69.9–75.9)	0.660 (0.651–0.669)	0.661 (0.613–0.695)
Less experienced observers ($n = 3$)	79.1 (75.9–84.3)	0.726 (0.717–0.734)	0.727 (0.685–0.797)

Data are given as mean (range) for pairwise agreement and pairwise Cohen's kappa, and point estimate (95% CI) for Fleiss kappa.

Table 3 Interobserver agreement in 83 cases of adnexal mass with regard to papillary projections among seven observers before and after a consensus meeting

Parameter	*Agreement (%)*	*Fleiss kappa*	*Cohen's kappa*
First evaluation before meeting			
All observers ($n = 7$)	86.6 (75.9–92.3)	0.536 (0.522–0.550)	0.520 (0.148–0.747)
Experienced observers ($n = 4$)	86.1 (80.7–89.2)	0.579 (0.562–0.597)	0.578 (0.381–0.643)
Less experienced observers ($n = 3$)	87.1 (84.3–89.2)	0.441 (0.419–0.463)	0.434 (0.348–0.577)
First evaluation after meeting			
All observers ($n = 7$)	82.7 (68.7–92.8)	0.487 (0.474–0.500)	0.480 (0.194–0.787)
Experienced observers ($n = 4$)	83.9 (75.9–92.7)	0.588 (0.572–0.603)	0.602 (0.490–0.787)
Less experienced observers ($n = 3$)	84.7 (80.7–90.4)	0.397 (0.373–0.421)	0.414 (0.318–0.470)

Data are given as mean (range) for pairwise agreement and pairwise Cohen's kappa, and point estimate (95% CI) for Fleiss kappa.

both before and after the consensus meeting, with no clear improvement after the meeting. The average percentage agreement was 86.6% (Fleiss kappa, 0.536) before the meeting and 82.7% (Fleiss kappa, 0.487) after it. Pairwise agreement was highly variable (Cohen's kappa, 0.148–0.787). Both before and after the consensus meeting interobserver agreement beyond chance between the four experienced observers was better than that between the three less experienced observers. No substantial change in interobserver agreement was noted after the consensus meeting for either the experienced or the less experienced observers.

Intraobserver repeatability with regard to tumor type is shown in Table 4. For all observers intraobserver repeatability was very good both before and after the consensus meeting. The percentage agreement varied between 87% and 95% (kappa, 0.83–0.94) before the meeting and between 87% and 93% after it (kappa, 0.83–0.90). Intraobserver repeatability tended to be slightly better for the less experienced observers, but for all observers it was similar before and after the consensus meeting.

Intraobserver repeatability with regard to the presence of papillary projections is shown in Table 5. For all observers, intraobserver repeatability with regard to the presence of papillary projections was good or very good both before and after the consensus meeting, with no substantial difference between the less experienced and most experienced observers. The percentage agreement varied between 92% and 100% (kappa, 0.72–1.0) before the meeting and between 88% and 96% afterwards (kappa, 0.64–0.90).

During the consensus meeting it became clear that some of the differences between and within observers could be explained by the observers feeling uncertain about how to classify the gray-scale ultrasound features of the tumors, or by them having different opinions on how to classify them. The following questions arose:

1. How does one discriminate between a papillary projection and another solid component? This question was the one that caused most uncertainty, and it resulted in lengthy discussions. According to the IOTA terms and definitions paper a papillary projection is 'any solid projection into the cyst cavity from the cyst wall with a height of 3 mm or more'[9]. In some cases it is difficult to determine whether a solid part is protruding. Should the angle between the solid component and the cyst wall or septum be taken into account when defining a papillary projection? For example, should an acute angle define the solid part as a papillary projection and an angle of 90° or more define it as a solid part (Figures 1 and 2)? Can a papillary projection have any size as long as its height is ≥ 3 mm[16]? If not, at what size does a projection of solid tissue become a solid component rather than a papillary projection?
2. How does one distinguish a solid component in the periphery of an ovarian lesion from ovarian stroma not belonging to the mass (ovarian crescent sign[17]) (Figure 2)?
3. How does one distinguish a solid component from a conglomerate of densely packed septa or densely

Table 4 Intraobserver repeatability in 83 cases of adnexal mass before and after a consensus meeting when classifying tumors as unilocular, unilocular solid, multilocular, multilocular solid or solid

	Before meeting		*After meeting*	
Observer	*Agreement (%)*	*Cohen's kappa (95% CI)*	*Agreement (%)*	*Cohen's kappa (95% CI)*
Most experienced observers				
A	87.9	0.84 (0.75–0.93)	91.6	0.89 (0.81–0.97)
B	86.7	0.83 (0.74–0.92)	90.3	0.86 (0.78–0.94)
C	89.1	0.86 (0.78–0.94)	86.7	0.83 (0.73–0.92)
D	86.7	0.83 (0.74–0.92)	86.8	0.83 (0.74–0.92)
Less experienced observers				
E	92.7	0.90 (0.82–0.98)	92.7	0.90 (0.86–0.94)
F	95.1	0.94 (0.88–1.0)	90.3	0.87 (0.79–0.95)
G	95.2	0.94 (0.88–1.0)	86.8	0.83 (0.74–0.92)

Table 5 Intraobserver repeatability in 83 cases of adnexal mass with regard to the presence of papillary projections before and after a consensus meeting

	Before meeting		*After meeting*	
Observer	*Agreement (%)*	*Cohen's kappa (95% CI)*	*Agreement (%)*	*Cohen's kappa (95% CI)*
Most experienced observers				
A	91.6	0.72 (0.52–0.92)	96.4	0.89 (0.77–1.0)
B	94.0	0.83 (0.69–0.97)	95.1	0.90 (0.80–1.0)
C	92.8	0.77 (0.59–0.95)	88.0	0.67 (0.48–0.86)
D	94.0	0.80 (0.63–0.97)	92.8	0.81 (0.67–0.95)
Less experienced observers				
E	98.8	0.88 (0.66–1.0)	96.4	0.75 (0.47–1.0)
F	91.6	0.76 (0.59–0.93)	90.4	0.76 (0.60–0.92)
G	100.0	1.0 (1.0–1.0)	90.3	0.64 (0.41–0.87)

packed very small cysts (Figure 3)? How does one distinguish a solid component from a tangentially cut septum?

4. When determining whether a tumor is unilocular, unilocular solid, multilocular, multilocular solid or solid, should one also take into account color/power Doppler information? This question arose because the color/power Doppler clips evaluated after the gray-scale ultrasound clips sometimes revealed information that would have changed the classification of the lesion had it been taken into account. For example, in some cases a septum will be seen only in color Doppler mode, and color Doppler ultrasound may clarify whether one is dealing with a solid component or a blood clot or other amorphous tissue.
5. Should solid protrusions into the lumen of a tube – the lesion being recognized as a tube on the basis of pattern recognition – be classified as papillary projections if they have a height of 3 mm or more (Figure 4)[18,19]? Should a lesion (judged to be a tube on the basis of pattern recognition) be described as a unilocular lesion (because in real life it is most likely to be a unilocular lesion), despite complete septa being present, so that according to the IOTA criteria it is a multilocular lesion[18,19]?

After discussion the participants in the meeting reached consensus that when performing evaluations 3 and 4 of the clips the following rules should apply:

1. Neither the size of a solid component protruding into a cyst cavity (as long as its height is 3 mm or more) nor the angle between a solid component and the cyst wall or a septum should be taken into account when defining a papillary projection. If there is uncertainty about whether a solid component protrudes into a cyst cavity, i.e. whether it is a papillary projection or not, it should be classified as a papillary projection (worst-case scenario) (Figures 1 and 2).
2. If a lesion can be schematically described as two ellipses, one inside the other (Figure 5), the solid component should be described as ovarian tissue provided that its gray-scale ultrasound morphology is compatible with ovarian stroma (ovarian crescent sign[17]). If there is uncertainty about whether solid tissue represents ovarian tissue or a solid component of a mass, it should be classified as a solid component (worst-case scenario).
3. If there is uncertainty about a structure being solid tissue or a conglomerate of densely packed septa, densely packed small cysts or a tangentially cut septum, the structure should be classified as a solid component (worst-case scenario).
4. Masses should be classified as unilocular, unilocular solid, multilocular, multilocular solid or solid only on the basis of the gray-scale ultrasound image. Doppler findings should not be taken into account (this is how it was done in all evaluations).

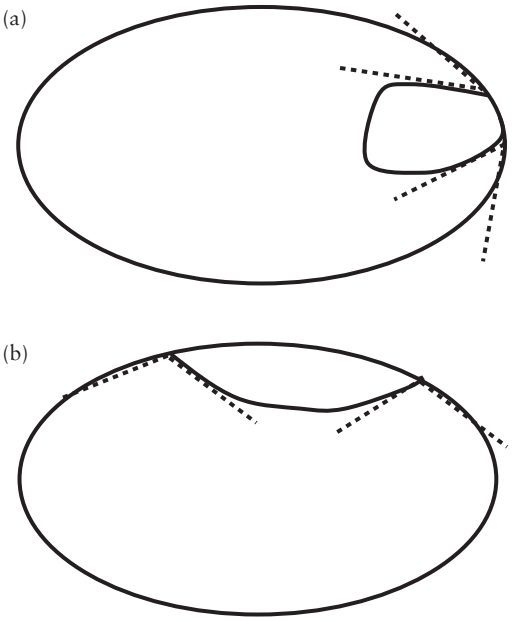

Figure 1 Schematic diagrams showing source of confusion among some observers as to whether a protrusion of solid tissue into a septum should be regarded as a papillary projection: (a) only if the angle (dotted lines) between the protrusion and the cyst wall or septum was acute (≤ 90°) or (b) also if it was obtuse (>90°). Consensus was reached that when performing analyses 3 and 4 the angle should not be taken into account, but any protrusion should be regarded as a papillary projection.

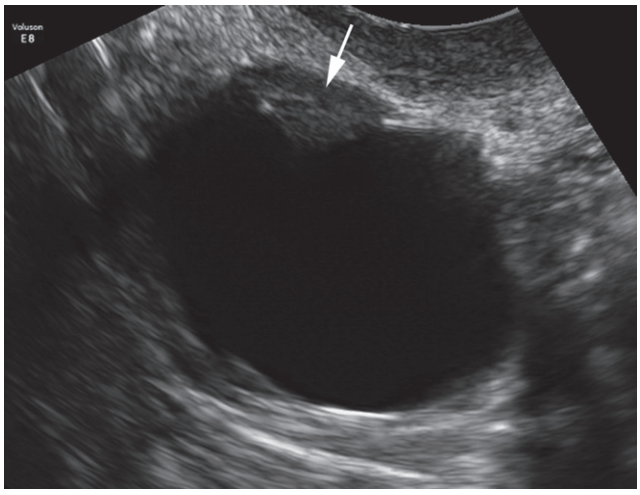

Figure 2 Ultrasound image illustrating difficulty in ascertaining whether a solid component (arrow) is part of the lesion or represents ovarian stroma, and whether or not, in this case, the protruding tissue should be regarded as a papillary projection. Consensus was reached that when performing analyses 3 and 4 if there is uncertainty about whether solid tissue represents ovarian tissue or a solid component of a mass, it should be classified as a solid component of the mass (worst-case scenario).

5. One should describe what one sees and not interpret the image using pattern recognition. If according to pattern recognition a mass is a diseased tube (which in all likelihood is a unilocular structure), but the ultrasound image reveals a multilocular structure with complete septa, the lesion should be classified as a multilocular cyst. If the 'cogwheel' sign or 'beads-on-a-string' sign is present, the solid protrusions should be classified as papillary projections if their height is at least 3 mm[20].

DISCUSSION

We found intraobserver repeatability to be very good for tumor type and good for papillary projections, and interobserver agreement to be good for tumor type. Interobserver agreement with regard to papillary projections, however, was only moderate – and highly variable between observer pairs – with no unequivocal improvement after the observers had discussed how to define a papillary projection at the consensus meeting. Moreover, we identified ultrasound images associated with uncertainty or disagreement between observers as to how they should be described using IOTA terminology. The definition of papillary projection caused most disagreement and confusion, but there was also uncertainty and disagreement about whether a certain structure should be classified as a solid component or as a conglomerate of septa, a conglomerate of small cysts or as ovarian stroma.

The strength of our study is that it provides information in an area that is poorly elucidated in the literature. Intraobserver repeatability and interobserver agreement with regard to tumor type and papillary projections have been described for only one observer pair[10], and to the best of our knowledge, problems that ultrasound examiners may encounter when they try to describe gray-scale ultrasound images of adnexal masses using the IOTA terminology have not been reported previously.

Limitations of our study are that we did not perform any sample size calculations and that the confidence limits for the kappa values relating to intraobserver agreement with regard to papillary projections were wide. Another limitation is that we used digital clips to estimate repeatability and agreement. Our results might not be applicable to live ultrasound examinations, in which scanning conditions may change from one minute to another and changes in settings may affect the image, and in which the interactive nature of a live scan may facilitate image interpretation. This could result in either poorer or better repeatability/agreement when live scanning is used than when digital clips are used. On the other hand, using digital clips means that all observers were exposed to exactly the same images, so any differences within or between observers are explained exclusively by differences in image interpretation. In any case, it would not have been possible to submit each patient to ultrasound examination by seven sonologists, each performing two examinations.

A possible source of bias in our study is that observers A, B and C collected the clips. However, even if they had remembered the patients when they analyzed the clips 1 month to 5 years after they collected them (which is implausible), we find it highly unlikely that this would

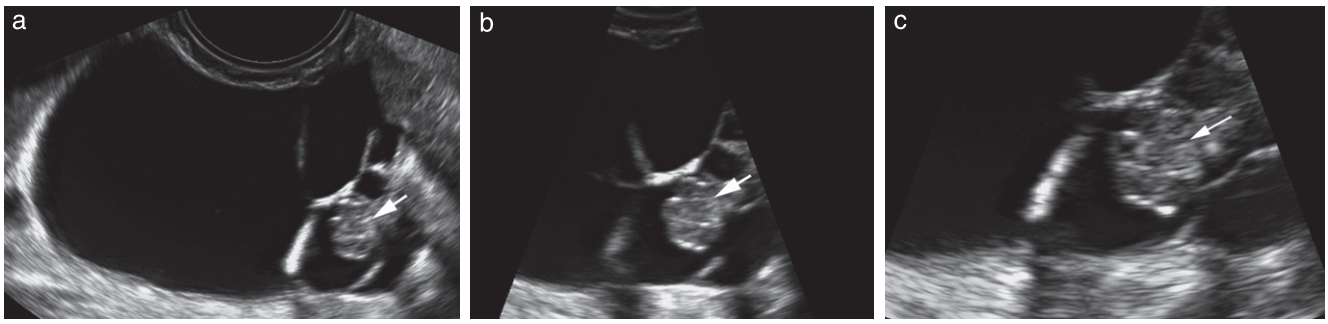

Figure 3 Ultrasound images, of the same adnexal mass, with increasing degree of magnification from (a) to (c) illustrating the difficulty in distinguishing a conglomerate of densely packed very small cysts from a solid component, in this case a papillary projection. The questionable structure is indicated with an arrow. Consensus was reached that when performing analyses 3 and 4 if there is uncertainty about a structure being solid tissue or a conglomerate of small cysts or densely packed septa, the structure should be classified as a solid component (worst-case scenario).

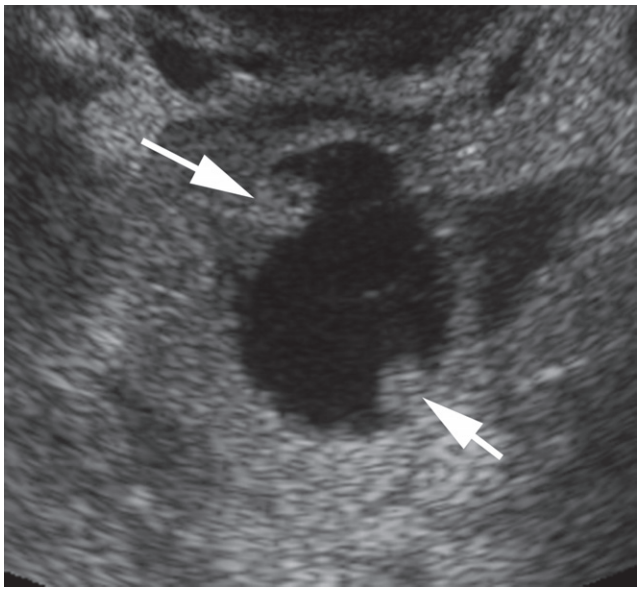

Figure 4 Ultrasound image showing a transverse section through a Fallopian tube with the 'cog-wheel' sign[18]. The lesion was recognized as a tube on the basis of pattern recognition. There was uncertainty among observers as to whether protrusions of solid tissue into the lumen of a tube (arrows) should be classified as papillary projections if they had a height of 3 mm or more. Consensus was reached that when performing analyses 3 and 4 one should describe what one sees and not interpret the image using pattern recognition: if the cogwheel or 'beads-on-a-string' sign[18] is present, solid protrusions should be classified as papillary projections if their height is at least 3 mm.

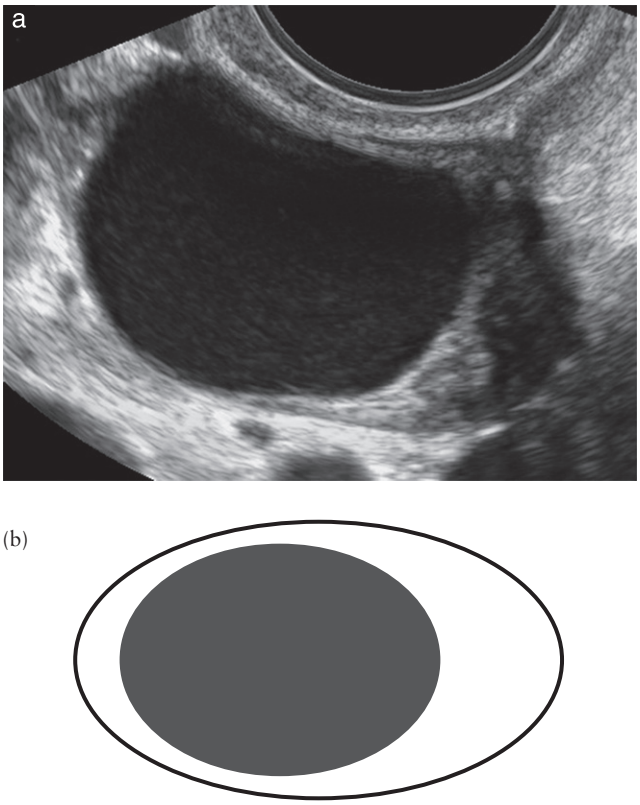

Figure 5 (a) Ultrasound image showing a longitudinal section through a mass and (b) schematic drawing corresponding to the image. Consensus was reached that this mass should be classified as a unilocular cyst and not as a unilocular solid mass, the solid component constituting ovarian stroma and not being part of the ovarian lesion. This is because if a lesion can be described schematically as two ellipses, one inside the other (b), the solid component surrounding the central cyst contents should be described as ovarian tissue, provided that its gray-scale ultrasound morphology is compatible with that of ovarian stroma.

have affected their classification of the masses with regard to tumor type or presence of papillary projections.

Our results are generalizable only to adnexal masses similar to those seen in our study. Had our study sample included a larger proportion of, or consisted exclusively of, tumors with ultrasound images difficult to interpret (with regard to the presence of solid components or papillary projections), intraobserver repeatability and interobserver agreement would almost certainly have been poorer. Moreover, our results are generalizable only to observers with ultrasound experience similar to the levels in our study. For observers with very limited ultrasound experience, intra-and interobserver agreement would probably be poorer than in this study.

Intraobserver repeatability and interobserver agreement with regard to tumor type and papillary projections have previously been reported for one observer pair (pair A–C in this study) analyzing three-dimensional ultrasound

volumes of adnexal masses[10]. The results were very similar to those reported here for observer pair A–C.

Our interpretation of the excellent intraobserver repeatability with regard to papillary projections is that each observer had his/her own definition of papillary projection and stuck to that definition when performing all four analyses of the digital clips. Each observer having his/her own definition of papillary projection also explains the very substantial variability in interobserver agreement with regard to the presence of papillary projections. The explanation for there being no unequivocal improvement in interobserver agreement with regard to papillary projections after the consensus meeting is that, despite lengthy discussions, the definition of papillary projection was not changed at the meeting, i.e. a papillary projection was defined as any protrusion of solid tissue into the cyst lumen but without a precise definition of the term protrusion. There was also no general improvement in intraobserver repeatability and interobserver agreement with regard to tumor type after the meeting. The most likely explanation for this is that there were few masses for which the observers had difficulty in interpreting the images with regard to the presence of solid components.

Solid components and papillary projections are variables in the IOTA logistic regression models LR1 and LR2[7]. Different interpretations with regard to the presence or absence of these variables affect the calculated risk of malignancy using LR1 and LR2. It is important to emphasize that the consensus reached between the observers at the consensus meeting cannot be seen as representative of a general consensus on how to define a papillary projection or other solid component of an adnexal mass. We do believe, however, that more precise definitions of papillary projections and solid components are needed, as well as practical advice on how to discriminate between, for example, a solid component and ovarian tissue, a solid component and a conglomerate of densely packed septa or small cysts, and a solid component and a tangentially cut septum.

It is difficult to predict the effect of more precise definitions on the diagnostic performance of the IOTA logistic regression models LR1 and LR2. Any effect on diagnostic performance with regard to classifying adnexal masses as benign or malignant would probably be small, because in a general population of adnexal masses most are easy to classify. However, more precise definitions would, we hope, result in risk estimates being more reproducible, which would be important should one want to use the absolute risk estimate clinically[10].

Even though we did not test the reproducibility of 'irregular cyst wall' in our study (irregular cyst wall is a variable included in both LR1 and LR2), we think that this is another variable that needs to be more precisely defined (for example, by providing representative images). This is because evaluation of irregularity is totally subjective and because the term 'irregular' is also used in the IOTA simple rules that have now been recommended for clinical use by the Royal College of Obstetricians and Gynaecologists in the UK[21–23].

ACKNOWLEDGMENTS

This work was supported by funds administered by two Swedish governmental grants: ALF-medel and Landstingsfinansierad Regional Forskning.

REFERENCES

1. Alcazar JL, Olartecoechea B, Guerriero S, Jurado M. Expectant management of adnexal masses in selected premenopausal women: a prospective observational study. *Ultrasound Obstet Gynecol* 2013; **41**: 582–588.
2. Medeiros LR, Fachel JM, Garry R, Stein AT, Furness S. Laparoscopy versus laparotomy for benign ovarian tumours. *Cochrane Database Syst Rev* 2005 (**3**): CD004751.
3. Carley ME, Klingele CJ, Gebhart JB, Webb MJ, Wilson TO. Laparoscopy versus laparotomy in the management of benign unilateral adnexal masses. *J Am Assoc Gynecol Laparosc* 2002; **9**: 321–326.
4. Valentin L, Hagen B, Tingulstad S, Eik-Nes S. Comparison of 'pattern recognition' and logistic regression models for discrimination between benign and malignant pelvic masses: a prospective cross validation. *Ultrasound Obstet Gynecol* 2001; **18**: 357–365.
5. 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.
6. Van Holsbeke C, Van Calster B, Bourne T, Ajossa S, Testa AC, Guerriero S, Fruscio R, Lissoni AA, Czekierdowski A, Savelli L, Van Huffel S, Valentin L, Timmerman D. External validation of diagnostic models to estimate the risk of malignancy in adnexal masses. *Clin Cancer Res* 2012; **18**: 815–825.
7. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L. 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.
8. Kaijser J, Bourne T, Valentin L, Sayasneh A, Van Holsbeke C, Vergote I, Testa AC, Franchi D, Van Calster B, Timmerman D. Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. *Ultrasound Obstet Gynecol* 2013; **41**: 9–20.
9. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. 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.
10. 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.
11. 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.
12. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; **76**: 378–382.
13. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measurement* 1960; **20**: 37–46.
14. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; **228**: 303–308.
15. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *BMJ* 1992; **304**: 1491–1494.
16. Hassen K, Ghossain MA, Rousset P, Sciot C, Hugol D, Baddoura R, Vadrot D, Buy JN. Characterization of papillary

projections in benign versus borderline and malignant ovarian masses on conventional and color Doppler ultrasound. *AJR Am J Roentgenol* 2011; **196**: 1444–1449.
17. Van Holsbeke C, Van Belle V, Leone FP, Guerriero S, Paladini D, Melis GB, Greggi S, Fischerova D, De Jonge E, Neven P, Bourne T, Valentin L, Van Huffel S, Timmerman D. Prospective external validation of the 'ovarian crescent sign' as a single ultrasound parameter to distinguish between benign and malignant adnexal pathology. *Ultrasound Obstet Gynecol* 2010; **36**: 81–87.
18. Valentin L. Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. *Ultrasound Obstet Gynecol* 1999; **14**: 338–347.
19. Valentin L. Use of morphology to characterize and manage common adnexal masses. *Best Pract Res Clin Obstet Gynaecol* 2004; **18**: 71–89.
20. Timor-Tritsch IE, Lerner JP, Monteagudo A, Murphy KE, Heller DS. Transvaginal sonographic markers of tubal inflammatory disease. *Ultrasound Obstet Gynecol* 1998; **12**: 56–66.
21. 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.
22. 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.
23. RCOG.org: Management of Suspected Ovarian Masses in Premenopausal Women (Green-top 62). http://www.rcog.org.uk/womens-health/clinical-guidance/ovarian-masses-premenopausal-women-management-suspected-green-top-62 (accessed May 16 2013).