Original Article

Gynecologic and Obstetric Investigation

Gynecol Obstet Invest 2010;69:160–168
DOI: 10.1159/000265012

Received: April 29, 2009
Accepted after revision: October 12, 2009
Published online: December 11, 2009

Ultrasound Experience Substantially Impacts on Diagnostic Performance and Confidence when Adnexal Masses Are Classified Using Pattern Recognition

Caroline Van Holsbeke[a, b] Anneleen Daemen[c] Joseph Yazbek[d] Tom K. Holland[d] Tom Bourne[a, e] Tinne Mesens[b] Lore Lannoo[a] Anne-Sophie Boes[a] Annelies Joos[a] Arne Van De Vijver[a] Nele Roggen[a] Bart de Moor[c] Eric de Jonge[b] Antonia C. Testa[f] Lil Valentin[g] Davor Jurkovic[d] Dirk Timmerman[a]

Department of Obstetrics and Gynaecology, [a]University Hospitals Leuven, and [b]Ziekenhuis Oost-Limburg, Genk, [c]Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Belgium; [d]Early Pregnancy and Gynaecology Assessment Unit, Kings' College Hospital, [e]Early Pregnancy and Gynaecological Ultrasound Unit, Imperial College Hammersmith Campus, London, UK; [f]Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Rome, Italy; [g]Department of Obstetrics and Gynaecology, Malmö University Hospital, Lund University, Malmö, Sweden

Key Words

Ultrasound experience · Diagnostic confidence · Ovarian tumors · Subjective impression · Pattern recognition · Risk of malignancy · Statistical models

Abstract

Aim: To determine how accurately and confidently examiners with different levels of ultrasound experience can classify adnexal masses as benign or malignant and suggest a specific histological diagnosis when evaluating ultrasound images using pattern recognition. ***Methods:*** Ultrasound images of selected adnexal masses were evaluated by 3 expert sonologists, 2 senior and 4 junior trainees. They were instructed to classify the masses using pattern recognition as benign or malignant, to state the level of confidence with which this classification was made and to suggest a specific histological diagnosis. Sensitivity, specificity, accuracy and positive and negative likelihood ratios (LR+ and LR–) with regard to malignancy were calculated. The area under the receiver operating characteristic curve (AUC) of pattern recognition was calculated by using six levels of diagnostic confidence. ***Results:*** 166 masses were examined, of which 42% were malignant. Sensitivity with regard to malignancy ranged from 80 to 86% for the experts, was 70 and 84% for the 2 senior trainees and ranged from 70 to 86% for the junior trainees. The specificity of the experts ranged from 79 to 91%, was 77 and 89% for the senior trainees and ranged from 59 to 83% for the junior trainees. The experts were uncertain about their diagnosis in 4–13% of the cases, the senior trainees in 15–20% and the junior trainees in 67–100% of the cases. The AUCs ranged from 0.861 to 0.922 for the experts, were 0.842 and 0.855 for the senior trainees, and ranged from 0.726 to 0.795 for the junior trainees. The experts suggested a correct specific histological diagnosis in 69–77% of the cases. All 6 trainees did so significantly less

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0378–7346/10/0693–0160$26.00/0

Accessible online at:
www.karger.com/goi

Caroline Van Holsbeke
Department of Obstetrics and Gynaecology, Z.O.L. Genk, Belgium
Schiepse Bos 6
BE–3600 Genk (Belgium)
Tel. +32 89 327 547, Fax +32 89 327 920, E-Mail caroline.van.holsbeke@skynet.be

often (22–42% of the cases). ***Conclusion:*** Expert sonologists can accurately classify adnexal masses as benign or malignant and can successfully predict the specific histological diagnosis in many cases. Whilst less experienced operators perform reasonably well when predicting the benign or malignant nature of the mass, they do so with a very low level of diagnostic confidence and are unable to state the likely histology of a mass in most cases.

Introduction

Several reports have demonstrated that subjective evaluation by expert sonologists (pattern recognition) is superior to the use of scoring systems and mathematical models when classifying adnexal masses as benign or malignant [1–3]. Only one study has assessed the results of subjective evaluation by less experienced examiners [4]. In the latter study, images of a consecutive series of 300 adnexal masses were evaluated. Timmerman et al. [4] showed that the test sensitivity in the hands of 2 expert sonologists was 96 and 98% and the specificity 90 and 89%, while the sensitivity and specificity for a moderately experienced examiner were 82 and 92%. The sensitivity of 3 inexperienced sonologists ranged from 87 to 90%, and the specificity from 81 to 85%.

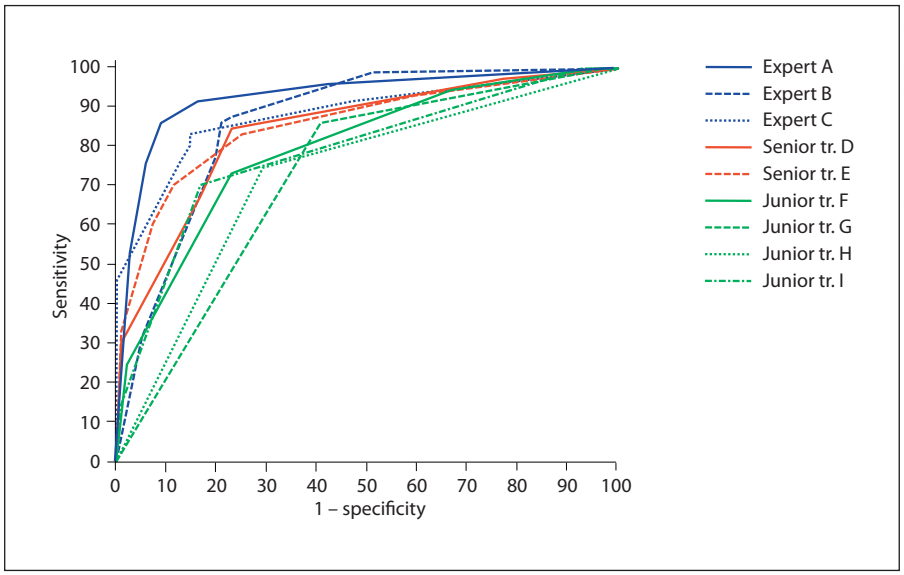
The aim of this study was to evaluate how accurately and confidently examiners with different levels of ultrasound experience can classify adnexal masses as benign or malignant and suggest a specific histological diagnosis when evaluating static ultrasound images of the masses using pattern recognition.

Methods

The database of the Early Pregnancy and Gynaecology Assessment Unit at King's College Hospital, London, was searched to identify all women who were diagnosed with adnexal tumors in the period between January 2004 and June 2006. Only women who underwent surgery and in whom a final histological diagnosis was available were included.

The cases were selected arbitrarily to ensure that the dataset included a mix of representative examples of benign, borderline and invasive malignant ovarian tumors. The number of masses with obviously benign or malignant ultrasound morphology ('easy tumors') was restricted in order to get a selected dataset with a high proportion of difficult to classify lesions. All women had been examined preoperatively by an expert sonologist (D.J.) with more than 10 years' experience in gynecological ultrasonography. The masses were classified according to the World Health Organization guidelines for histology [5].

The study formed a part of the multicenter IOTA (International Ovarian Tumor Analysis) collaboration, which was approved by the local hospital ethics committees [6, 7]. Representative gray-scale and color Doppler images of the masses were made by an expert sonologist (D.J.), anonymized and saved on a hard disk. Color Doppler images were not available for all of the masses, but the written reports always contained information on the color score. The color score is a subjective score between 1 and 4 assigned by the sonologist and indicating the amount of detectable color Doppler signals (reflecting vascularization) inside a mass [6]. After the images had been anonymized, they were evaluated independently by 9 observers. The observers were blinded to each other, the results and to the histological diagnosis. They had access to relevant clinical information (indication for the scan, symptoms, palpable mass), information on personal and family history of ovarian cancer and information on the color score if color Doppler images were not available. For each mass, the observers noted their answer to the following three questions: (1) 'Based on your subjective impression, do you think this mass is benign or malignant?', (2) 'How confident are you about your benign or malignant classification: certainly benign, probably benign, uncertain (= complete uncertainty about the mass being benign or malignant), probably malignant or certainly malignant, and (3) 'Which specific histological diagnosis would you suggest?'. The observers could choose one of eleven predefined specific histological diagnoses: simple cyst/functional cyst, dermoid, endometrioma, cystadenofibroma, abscess, rare benign tumor, mucinous borderline tumor, serous borderline tumor, primary invasive tumor or rare malignant tumor. Mucinous borderline tumor, serous borderline tumor, primary invasive tumor and rare malignancy were regarded as specific diagnoses. The observers were 3 expert sonologists (L.V., A.T., D.T.), 2 senior trainees (T.M., L.L.), and 4 junior trainees (A.B., A.J., N.R., A.V.). The experts (further on referred to as experts A, B and C) were senior gynecologists in tertiary referral gynecologic ultrasound units and had each performed over 5,000 scans. According to the guidelines of the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB), they were level 3 practitioners [8]. All trainees were trainees in obstetrics and gynecology. The senior trainees (referred to as senior trainees D and E), who were both in their 5th year of training, were moderately experienced and had received at least 1 year of training in gynecologic ultrasound in the ultrasound department of one of the experts (D.T.), where they had carried out over 700 scans each (level 1 practitioners according to EFSUMB guidelines) [8]. The junior trainees (referred to as junior trainees F, G, H and I), who were at the start of their 1st year of training, had attended ultrasound lectures and gained basic knowledge on the ultrasound morphology of adnexal masses during undergraduate training but lacked formal practical ultrasound training. The performance of the less experienced observers was compared with the performance of the 'consensus opinion', the latter being defined as the ultrasound diagnosis suggested by at least 2 of the 3 expert observers. For six adnexal masses all 3 experts suggested a different histological diagnosis. In these cases, the histological diagnosis predicted by the expert sonologist who had performed the real-time ultrasound examination was considered to be the consensus opinion.



Color version available online

Fig. 1. Receiver operating characteristic curves for 9 sonologists using pattern recognition to classify static images of adnexal masses as benign or malignant. The sonologists represented different levels of ultrasound expertise: 3 were experts, 2 were senior trainees (senior tr.), and 4 were junior trainees (junior tr.).

Table 1. Histopathological diagnoses of the 166 masses

Histopathological diagnosis	n	%
Benign (n = 96; 57.8%)		
Dermoid	35	21.1
Cystadenoma/fibroma	35	21.1
Endometrioma	16	9.6
Fibroma	6	3.6
Simple cyst/functional cyst	2	1.2
Abscess	1	0.6
Rare benign tumor	1	0.6
Malignant (n = 70; 42.2%)		
Mucinous borderline	16	9.6
Serous borderline	18	10.8
Common invasive (epithelial)	25	15.1
Rare invasive (nonepithelial)[1]	11	6.6

[1] For example: dysgerminoma, yolk sac tumor and granulosa cell tumor.

Statistical Analysis

Statistical analyses were performed using SAS version 9.1.3 for Windows (SAS Institute Inc., Cary, N.C., USA, 2002–2003). The sensitivity, specificity and accuracy for the prediction of the character of an adnexal mass of the junior and senior trainees were compared with those of the consensus opinion, and the statistical significance of differences in sensitivity, specificity and accuracy was determined using McNemar's test. The diagnostic performance was also expressed as positive and negative likelihood ratios (LR+ and LR–). The 95% confidence intervals for accuracy, sensitivity and specificity were calculated with the Wilson's score interval method [9], and for LR+ and LR– they were calculated by the Cox-Hinkley-Miettinen-Nurminen method [10].

The area under the receiver operating characteristics curve (AUC) for pattern recognition was calculated using the six levels of diagnostic confidence as 'cut-off points' (certainly benign, probably benign, uncertain but in the dichotomous classification stated to be benign, uncertain but in the dichotomous classification stated to be malignant, probably malignant, and certainly malignant).

The statistical significance of differences in AUC was determined as described by DeLong et al. [11]. Two-tailed p values <0.05 were considered statistically significant.

Results

Of the 166 masses in the database, 70 (42%) were malignant. Table 1 shows the histopathological diagnoses.

The sensitivity, specificity, accuracy and positive and negative LRs with regard to malignancy predicted by the 9 observers are shown in table 2. For the 'consensus opinion', the sensitivity was 83%, the specificity 86%, the LR+ was 6.12, the LR– 0.20 and the accuracy 85%. The accuracy of all 4 junior trainees was significantly poorer than that of the 'consensus opinion' of the experts. The accuracy of the senior trainees was also lower than that of the 'consensus opinion', but the differences did not reach statistical significance. The experts were uncertain about their diagnosis (benign or malignant) in 4–13% of the cases, the senior trainees were uncertain in 15–20% of the cases and the junior trainees in 67–100% of the cases (table 3). The AUCs ranged from 0.861 to 0.922 for the experts, were 0.842 and 0.855 for the senior trainees and ranged from 0.726 to 0.795 for the junior trainees (table 2; fig. 1). The AUC of the

Table 2. Accuracy, sensitivity, specificity, positive and negative LR with regard to malignancy of subjective evaluation of static ultrasound images by observers with varying levels of ultrasound experience

Sonologist	AUC	Accuracy n (%)	95% CI	p	Sensitivity n (%)	95% CI	p	Specificity n (%)	95% CI	p	LR+ (95% CI)	LR– (95% CI)
Experts												
A	0.92247	89 (147/166)	83–93		86 (60/70)	76–92		91 (87/96)	83–95		9.14 (5.03–17.25)	0.16 (0.09–0.27)
B	0.86109	82 (136/166)	75–87		86 (60/70)	76–92		79 (76/96)	70–86		4.11 (2.81–6.23)	0.18 (0.10–0.31)
C	0.88199	83 (138/166)	77–88		80 (56/70)	69–88		85 (82/96)	77–91		5.49 (3.41–9.12)	0.23 (0.14–0.36)
Consensus opinion		85 (141/166)	79–90		83 (58/70)	72–90		86 (83/96)	78–92		6.12 (3.74–10.36)	0.20 (0.12–0.32)
Senior trainees												
D	0.84189	80 (133/166)	73–85	0.1441	84 (59/70)	74–91	0.7630	77 (74/96)	68–84	0.0389	3.68 (2.56–5.45)	0.20 (0.12–0.34)
E	0.85506	81 (134/166)	74–86	0.1779	70 (49/70)	58–79	0.0201	89 (85/96)	81–93	0.5637	6.11 (3.52–10.95)	0.34 (0.23–0.47)
Junior trainees												
F	0.78586	78 (129/166)	71–83	0.0455	70 (49/70)	58–79	0.0290	83 (80/96)	75–89	0.4913	4.20 (2.67–6.81)	0.36 (0.24–0.51)
G	0.72560	72 (120/166)	65–79	0.0014	74 (52/70)	63–83	0.1336	71 (68/96)	61–79	0.0039	2.55 (1.83–3.62)	0.36 (0.23–0.54)
H	0.72664	70 (117/166)	63–77	0.0004	86 (60/70)	76–92	0.6171	59 (57/96)	49–69	<0.0001	2.11 (1.65–2.77)	0.24 (0.12–0.47)
I	0.79464	75 (125/166)	68–81	0.0114	73 (51/70)	61–82	0.0896	77 (74/96)	68–84	0.0606	3.18 (2.18–4.77)	0.35 (0.23–0.51)

p value refers to the comparison with the consensus opinion. Consensus opinion is defined as the diagnosis suggested by at least 2 of the 3 experts.

Table 3. Diagnostic confidence with regard to malignancy of ultrasound observers with varying levels of experience

Sonologist	Diagnostic confidence								
	certain			probable			uncertain		
	overall	benign	malignant	overall	benign	malignant	overall	benign	malignant
Expert									
A	98 (59)	58	40	47 (28)	28	19	21 (13)	10	11
B	75 (45)	52	23	81 (49)	41	40	10 (6)	3	7
C	89 (54)	51	38	70 (42)	43	27	7 (4)	2	5
Senior trainee									
D	46 (28)	24	22	95 (57)	54	41	25 (15)	7	18
E	68 (41)	44	24	65 (39)	40	25	33 (20)	22	11
Junior trainee									
F	0	0	0	18 (11)	7	11	148 (89)	94	54
G	0	0	0	0	0	0	166 (100)	86	80
H	3 (2)	3	0	5 (3)	5	0	158 (95)	59	99
I	5 (3)	5	0	50 (30)	31	19	111 (67)	57	54

Figures indicate numbers of cases (%).

best expert (i.e. the expert that had the largest AUC) was significantly larger than the AUCs of all 6 trainees, and 2 of the junior trainees had AUCs that were significantly smaller than those of the senior trainees (table 4).

The diagnostic performance of pattern recognition for predicting a specific histological diagnosis is presented in tables 5–7. The experts suggested a correct specific diagnosis in 71–77% of the cases, both senior trainees in 42% of the cases ($p < 0.0001$ when comparing with the consensus opinion) and the junior trainees in 22–42% of the cases ($p < 0.0001$ when comparing with the consensus opinion). The benign histologies that were best classified by the experts were dermoid cysts (sensitivity between 77 and 91%, specificity between 95 and 97%) and endome-

Table 4. p values when comparing AUC of 9 observers using pattern recognition for classifying adnexal masses as benign or malignant

Sonologist	Senior trainee		Junior trainee			
	D	E	F	G	H	I
Expert A	0.0071	0.0241	<0.0001	<0.0001	<0.0001	0.0003
Expert B	0.5412	0.8537	0.0370	0.0002	0.0001	0.0636
Expert C	0.1753	0.3452	0.0037	0.0002	<0.0001	0.0091
Senior trainee D			0.1382	0.0052	0.0032	0.2072
Senior trainee E			0.0555	0.0014	0.0008	0.0802

AUC values were compared using the method of DeLong et al. [11].

triomas (sensitivity 88% for all 3 experts and specificity between 97 and 99%). The sensitivity with regard to the specific benign histological diagnoses of the 6 less experienced sonologists was low, especially with regard to dermoid cyst, and the sensitivity with regard to specific malignant diagnoses was also low. For 3 of the 4 junior trainees, both the sensitivity and specificity with regard to primary invasive malignancy were significantly lower than those of the consensus opinion.

Table 5. Percentage of correct specific histological diagnoses suggested by ultrasound observers with different levels of experience using pattern recognition

	Correct specific diagnosis, %	p value[1]
Experts		
A	77 (128/166)	
B	69 (115/166)	
C	71 (118/166)	
Consensus opinion[2]	77 (128/166)	
Senior trainees		
D	42 (89/166)	<0.0001
E	42 (71/166)	<0.0001
Junior trainees		
F	42 (70/166)	<0.0001
G	30 (50/166)	<0.0001
H	22 (37/166)	<0.0001
I	40 (67/166)	<0.0001

[1] Comparison with the consensus opinion.
[2] Diagnosis suggested by at least 2 of the 3 experts.

Discussion

We have demonstrated that expert sonologists can correctly discriminate between benign and malignant adnexal masses and make a correct specific histological diagnosis in most cases by evaluating static representative ultrasound images. More and more studies demonstrate that pattern recognition during ultrasound should be the gold standard in the preoperative assessment of adnexal masses, but one should take into account that less experienced examiners are much less capable of doing this [1–3, 12–14]. Even though the 2 senior trainees were surprisingly good in distinguishing benign from malignant adnexal masses, they classified the masses with so little confidence that the clinical value of their suggested diagnosis must be questioned. However, it is important to emphasize that the tumors in this study are selected, a high proportion of the tumors being borderline tumors, which are very difficult to classify as benign or malignant [15], and only a few being simple cysts or functional cysts. This means that a higher proportion of the tumors in this series than in an unselected tumor population seen in a gynecological outpatient clinic was difficult to classify.

The strength of our study is that we have compared not only the ability to distinguish between benign and malignant adnexal masses and the ability to make a correct histological diagnosis between observers with different levels of experience but also compared their diagnostic confidence. To the best of our knowledge, the issue of diagnostic confidence has not been addressed in any published study. The diagnostic confidence is important, because it is difficult to make correct clinical decisions when the suggested diagnosis is uncertain.

A limitation of our study is that pattern recognition was evaluated using static ultrasound images. In another study we have shown that whilst the sensitivity is similar, the specificity with regard to malignancy of a real-time ultrasound examination is higher than that derived from the evaluation of static ultrasound images alone [16].

However, there are major practical difficulties associated with performing a study where patients are examined by more than 2 sonologists. A second limitation of our study is that the representative ultrasound images were all taken by an expert sonologist and that in most cases the color score was provided by that expert. This may have led to an overestimation of the diagnostic performance of pattern recognition in the hands of the less experienced examiners, because they did not need to create the ultrasound images themselves, nor did they need to assign a color score. Had they been required to do so, their diagnostic performance is likely to have been poorer than it was in this study. The experts, however, might have done better had they themselves performed the ultrasound examinations. A third limitation of our study is that the tumor population was selected to include a high proportion of tumors that were not obviously benign or malignant. The reason for this is that differences in diagnostic performance of pattern recognition between individuals with varying levels of ultrasound experience might be easier to detect in a population containing a high proportion of difficult tumors. The differences between the experts and non-experts are likely to be smaller in a nonselected tumor population.

Our results are in agreement with those of Timmerman et al. [4] and those of Guerriero et al. [17], that the ability to correctly discriminate between benign and malignant adnexal masses when evaluating static ultrasound images using pattern recognition increases with the experience of the observers. Guerriero et al. [17] reported how accurately endometriomas, teratomas and serous cysts could be diagnosed when ultrasound images were independently evaluated by sonologists with different levels of experience. Their conclusion was the same as ours, i.e. the performance of sonologists improves with increasing level of experience and endometriomas are easier to diagnose than teratomas [18]. However, their 'non-experts' had more experience in gynecological ultrasound than our 'non-experts', and the performance of the experts was better in our study than in theirs. This is probably explained by the fact that in the study of Guerriero et al. [17] a mass could only be classified as belonging to a specific histological category if it fitted a predetermined definition (for example, an endometrioma was defined as 'round or ovoid homogeneous hypoechoic tissue with ground glass content without papillary proliferation and a clear demarcation from the ovarian parenchyma'). In our study, pattern recognition was used, giving the examiner more freedom to use his/her skills in subjective evaluation of ultrasound findings.

The classification of adnexal masses using pattern recognition by an expert sonologist is more accurate than any other method, e.g. scoring systems or classification systems [1, 2] and mathematical models [3]. However, as this study demonstrates when using pattern recognition the experience of the ultrasound examiner is important. Yazbek et al. [19] showed that the preoperative risk assessment of an adnexal mass based on the use of pattern recognition by a gynecologist with expertise in gynecological ultrasound resulted in fewer patients undergoing unnecessary staging laparotomies for benign pathology than if the patient had been scanned preoperatively by a non-expert and in fewer patients with a malignancy undergoing surgery by laparoscopy and/or by gynecologists not specialized in gynecologic oncology [19]. Mathematical models and scoring systems might help less experienced examiners achieve the same performance as pattern recognition by experts. A few models seem to perform as well as pattern recognition when they are used by experts in gynecological ultrasound [14 and unpublished IOTA data], but the performance of these mathematical models in the hands of non-experts remains to be determined.

In clinical practice, it is important not only to be able to distinguish between benign and malignant tumors but also to make a correct specific histological diagnosis. This is particularly true of premenopausal patients that want to preserve their fertility. In many centers, patients with endometriosis (endometrioma may be an indicator of deep infiltrating endometriosis) are referred to gynecologic surgeons with special expertise in this field. If a lesion is likely to be a peritoneal pseudocyst, a simple cyst or a small dermoid cyst, expectant management may be adopted. In this study, dermoid cysts were the type of cyst that was most often correctly classified by the expert sonologists (sensitivity of 91%, specificity 98%). The sensitivity of the less experienced sonologists was much lower (6 and 29%), even though their specificity was as high as that of the experts (97–99%). In other studies, sensitivities with regard to dermoid cysts ranged from 53 to 100% and specificity from 94 to 100% [20–25]. Published sensitivities (81–92%) and specificities (89–97%) with regard to endometrioma are similar to those of our experts but higher than those of the less experienced examiners in our study [20–22, 24–28].

Our work shows that experience is necessary for optimal use of pattern recognition when classifying adnexal masses. Because pattern recognition is the best method for making a correct diagnosis in an adnexal mass, it seems justified to spend time and resources on training gynecologists in using pattern recognition.

Table 6. Sensitivity and specificity of subjective evaluation of static ultrasound images for the prediction of specific benign histological diagnoses for observers with varying levels of ultrasound experience

Sonologist	Dermoid (n = 35)						Cystadenofibroma (n = 35)					
	sensitivity		p	specificity		p	sensitivity		p	specificity		p
	%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Experts												
A	91 (32/35)	78–97		97 (127/131)	92–98.8		74 (26/35)	58–86		89 (117/131)	83–94	
B	89 (31/35)	74–95		95 (124/131)	89–97		40 (14/35)	26–56		95 (124/131)	89–97	
C	77 (27/35)	61–88		97 (127/131)	92–98.8		74 (26/35)	58–86		86 (113/131)	79–91	
Consensus opinion	91 (32/35)	78–97		98 (128/131)	93–99.2		69 (24/35)	52–81		89 (117/131)	83–94	
Senior trainees												
D	23 (8/35)	12–39	<0.0001	98 (128/131)	93–99.2	1.0000	46 (16/35)	30–62	0.0455	89 (117/131)	83–94	1.0000
E	29 (10/35)	16–45	<0.0001	99 (130/131)	96–99.9	0.3173	63 (22/35)	46–77	0.5637	80 (105/131)	73–86	0.0143
Junior trainees												
F	14 (5/35)	6–29	<0.0001	98 (129/131)	95–99.6	0.6547	57 (20/35)	41–72	0.2482	80 (105/131)	73–86	0.0233
G	17 (6/35)	8–33	<0.0001	99 (130/131)	96–99.9	0.1573	26 (9/35)	14–42	0.0006	77 (101/131)	69–83	0.0094
H	6 (2/35)	2–19	<0.0001	99 (130/131)	96–99.9	0.3173	17 (6/35)	8–33	0.0001	93 (122/131)	87–96	0.2971
I	20 (7/35)	10–36	<0.0001	97 (127/131)	92–98.8	0.7055	60 (21/35)	44–74	0.3173	86 (113/131)	79–91	0.4142

p value when results are compared with those of the consensus opinion. Consensus opinion is defined as the diagnosis predicted by at least 2 of the 3 experts.

Table 7. Sensitivity and specificity of subjective evaluation of static ultrasound images for the prediction of specific malignant histological diagnoses of adnexal masses for observers with varying levels of ultrasound experience

Sonologist	Borderline malignant (n = 34)						Primary invasive (n = 25)					
	sensitivity		p	specificity		p	sensitivity		p	specificity		p
	%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Experts												
A	74 (25/34)	57–85		93 (123/132)	88–96		80 (20/25)	61–91		99 (140/141)	96–99.9	
B	71 (24/34)	54–83		87 (115/132)	80–92		80 (21/25)	65–94		94 (134/141)	90–98	
C	59 (20/34)	42–74		92 (122/132)	87–96		76 (19/25)	57–88		96 (136/141)	92–98	
Consensus opinion	68 (23/34)	51–81		91 (120/95)	85–95		88 (22/25)	70–96		99 (139/141)	95–99.6	
Senior trainees												
D	65 (22/34)	48–79	0.7630	81 (107/132)	74–87	0.0093	76 (19/25)	57–88	0.1797	94 (132/141)	88–97	0.0196
E	47 (16/34)	31–63	0.0348	83 (109/132)	75–88	0.0278	52 (13/25)	34–70	0.0027	98 (138/141)	94–99.3	0.6547
Junior trainees												
F	41 (14/34)	26–58	0.0067	87 (115/132)	80–92	0.2971	68 (17/25)	48–83	0.0588	96 (135/141)	91–98	0.1573
G	44 (15/34)	29–61	0.0325	81 (107/132)	74–87	0.0093	64 (16/25)	45–80	0.0143	84 (119/141)	78–89	<0.0001
H	65 (22/34)	48–79	0.7815	70 (92/132)	61–77	<0.0001	40 (10/25)	23–59	0.0027	84 (118/141)	77–89	<0.0001
I	56 (19/34)	39–71	0.2482	89 (117/132)	82–93	0.5485	72 (18/25)	52–86	0.0455	91 (128/141)	95–95	0.0045

p value when results are compared with those of the consensus opinion. Consensus opinion is defined as the diagnosis predicted by at least 2 of the 3 experts.

Endometrioma (n = 16)						Fibroma (n = 6)					
sensitivity		p	specificity		p	sensitivity		p	specificity		p
%	95% CI		%	95% CI		%	95% CI		%	95% CI	
88 (14/16)	64–96		99 (149/150)	96–99.9		50 (3/6)	19–81		100 (160/160)	98–100	
88 (14/16)	64–96		97 (146/150)	93–99		50 (3/6)	19–81		100 (160/160)	98–100	
88 (14/16)	64–96		99 (148/150)	95–99.6		50 (3/6)	19–81		100 (160/160)	98–100	
88 (14/16)	64–96		99 (149/150)	96–99.9		50 (3/6)	19–81		100 (160/160)	98–100	
69 (11/16)	44–86	0.1797	94 (141/150)	89–97	0.0114	17 (1/6)	3–56	0.1573	99 (158/160)	96–99.7	0.1573
56 (9/16)	33–77	0.0588	93 (139/150)	87–96	0.0039	17 (1/6)	3–56	0.3173	99 (159/160)	97–99.9	0.3174
69 (11/16)	44–86	0.1797	96 (144/150)	92–98	0.0588	33 (2/6)	10–70	0.5637	99 (158/160)	96–99.7	0.1573
6 (1/16)	1–28	0.0003	97 (146/150)	93–99	0.1797	33 (2/6)	10–70	0.5637	93 (149/160)	88–96	0.0009
19 (3/16)	7–43	0.0023	95 (143/150)	91–98	0.0339	0 (0/6)	0–39	0.0833	97 (155/160)	93–99	0.0254
50 (8/16)	28–72	0.0143	91 (136/150)	85–94	0.0008	17 (1/6)	3–56	0.3173	98 (157/160)	95–99.4	0.0833

Rare malignant (n = 11)					
sensitivity		p	specificity		p
%	95% CI		%	95% CI	
91 (10/11)	62–98		97 (151/155)	94–99	
73 (8/11)	43–90		98 (152/155)	94–99.3	
73 (8/11)	43–90		95 (148/155)	91–98	
82 (9/11)	52–95		98 (152/155)	94–99.3	
36 (4/11)	15–65	0.0253	99 (153/155)	95–99.6	0.6547
45 (5/11)	21–72	0.0455	100 (155/155)	98–100	0.0833
45 (5/11)	21–72	0.0455	96 (149/155)	92–98	0.3173
0 (0/11)	0–26	0.0027	99 (153/155)	95–99.6	0.6547
0 (0/11)	0–26	0.0027	97 (151/155)	94–99	0.6547
45 (5/11)	21–72	0.1025	98 (152/155)	94–99.3	1.0000

Acknowledgements

This work was supported by the Swedish Medical Research Council (grant No. K2006-73X-11605-11-3), funds administered by Malmö University Hospital, Allmänna Sjukhusets i Malmö Stiftelse för Bekämpande av Cancer (the Malmö General Hospital Foundation for Fighting against Cancer), Landstingsfinansierad regional forskning and ALF-medel (i.e. two Swedish governmental grants from the region of Scania).

The study was also supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, the Flemish Government: FWO projects G.0241.04, G.0499.04, G.0232.05, G.0318.05, G.0553.06, G.0302.07, research communities (ICCoS, ANMMM, MLDM); IWT: GBOU-McKnow-E, GBOU-ANA, TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3 and the Belgian Federal Science Policy Office: IUAP P6/25; EU-RTD: ERNSI; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

References

1 Valentin L, Jurkovic D, Van Calster B, Testa AC, Van Holsbeke C, Bourne T, Vergote I, Van Huffel S, Timmerman D: Adding a single CA-125 measurement to ultrasound performed by an experienced examiner does not improve preoperative discrimination between benign and malignant adnexal masses. A prospective international multicentre study of 809 patients. Ultrasound Obstet Gynecol 2009, in press.
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 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.
4 Timmerman D, Schwarzler P, Collins WP, Claerhout F, Coenen M, Amant F, et al: Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. Ultrasound Obstet Gynecol 1999;13:11–16.
5 Scully RE: World Health Organization. Histological Typing of Ovarian Tumours. Berlin, Springer, 1999.
6 Timmerman D, Valentin L, Bourne TH, et al: Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. Ultrasound Obstet Gynecol 2000;16:500–505.
7 Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML: 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.
8 EFSUMB Newsletter: Minimum training recommendations for the practice of medical ultrasound. Ultraschall Med 2006;26:84–86.
9 Newcombe RG: Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med 1998;17:857–872.
10 Miettinen OS, Nurminen M: Comparative analysis of two rates. Stat Med 1985;4:213–226.
11 DeLong E, DeLong D, Clarke-Pearson D: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845.
12 Van Calster B, Timmerman D, Borune T, Testa AC, Van Holsbeke C, Domali E, Jurkovic D, Neven P, Van Huffel S, Valentin L: Discrimination between benign and malignant adnexal masses by specialised ultrasound examination versus serum CA-125. J Natl Cancer Inst 2007;99:1706–1714.
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;30:257–281.
14 Van Holsbeke C, Van Calster B, Testa AC, Domali E, Lu C, Van Huffel S, Valentin L, Timmerman D: Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. Clin Cancer Res 2009;15:684–691.
15 Valentin L, Ameye L, Jurkovic D, Metzger U, Lécuru F, Van Huffel S, Timmerman D: Which extrauterine pelvic masses are difficult to correctly classify as benign or malignant on the basis of ultrasound findings and is there a way of making a correct diagnosis? Ultrasound Obstet Gynecol 2006;27:438–444.
16 Van Holsbeke C, Yazbek J, Holland TK, Daemen A, De Moor B, Testa AC, Valentin L, Jurkovic D, Timmerman D: Real-time ultrasound versus evaluation of static images in the preoperative evaluation of adnexal masses. Ultrasound Obstet Gynecol 2008;32:828–831.
17 Guerriero S, Alcazar JL, Pascual MA, Ajossa S, Gerada M, Bargellini R, Virgilio B, Melis G: Intraobserver and interobserver agreement of grayscale typical ultrasonographic patterns for the diagnosis of ovarian cancer. Ultrasound Med Biol 2008;34:1711–1716.
18 Guerriero S, Alcazar JL, Pascual MA, Ajossa S, Gerada M, Bargellini R, Virgilio B, Melis G: Diagnosis of the most frequent benign ovarian cysts: is ultrasonography accurate and reproducible? J Womens Health 2009;18:519–527.
19 Yazbek J, Raju SK, Ben-Nagi J, Holland TK, Hillaby K, Jurkovic D: Effect of quality of gynaecological ultrasonography on management of patients with suspected ovarian cancer: a randomised controlled trial. Lancet Oncol 2008;9:124–131.
20 Valentin L: Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. Ultrasound Obstet Gynecol 1999;14:338–347.
21 Benacerraf BR, Finkler NJ, Wojciechowski C, Knapp RC: Sonographic accuracy in the diagnosis of ovarian masses. J Reprod Med 1990;35:491–495.
22 Fleisher AC, James AE, Millis JB, Julian C: Differential diagnosis of pelvic masses by gray scale sonography. Am J Roentgenol 1978;131:469–476.
23 Mais V, Guerriero S, Ajossa S, Angiolucci M, Paoletti AM, Melis GB: Transvaginal sonography in the diagnosis of cystic teratoma. Obstet Gynecol 1995;85:48–52.
24 Guerriero S, Mallarini G, Ajossa A, Risalvato A, Satta R, Mais V, Angiolucci M, Melis GB: Transvaginal ultrasound and computed tomography combined with clinical parameters and CA-125 determinations in the differential diagnosis of persistent ovarian cysts in premenopausal women. Ultrasound Obstet Gynecol 1997;9:339–343.
25 Jermy K, Luise C, Bourne T: The characterization of common ovarian cysts in premenopausal women. Ultrasound Obstet Gynecol 2001;17:140–144.
26 Guerriero S, Mais V, Ajossa S, Paoletti AM, Angiolucci M, Labate F, Melis GB: The role of endovaginal ultrasound in differentiating endometriomas from other ovarian cysts. Clin Exp Obstet Gynecol 1995;22:20–22.
27 Guerriero S, Mais V, Ajossa S, Paoletti AM, Angiolucci M, Melis GB: Transvaginal ultrasonography combined with CA-125 plasma levels in the diagnosis of endometrioma. Fertil Steril 1996;65:293–298.
28 Mais V, Guerriero S, Ajossa S, Angiolucci M, Paoletti AM, Melis GB: The efficiency of transvaginal ultrasonography in the diagnosis of endometrioma. Fertil Steril 1993;60:776–780.