Ultrasound Obstet Gynecol 2009; **34**: 454–461
Published online 8 September 2009 in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/uog.6443

Ultrasound methods to distinguish between malignant and benign adnexal masses in the hands of examiners with different levels of experience

C. VAN HOLSBEKE*†, A. DAEMEN‡, J. YAZBEK§, T. K. HOLLAND§, T. BOURNE*¶, T. MESENS†, L. LANNOO*, B. DE MOOR‡, E. DE JONGE†, A. C. TESTA**, L. VALENTIN††, D. JURKOVIC§ and D. TIMMERMAN*

**Department of Obstetrics and Gynaecology, University Hospitals Leuven, †Department of Obstetrics and Gynaecology, Ziekenhuis Oost-Limburg, Genk and ‡Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Belgium, §Early Pregnancy and Gynaecology Assessment Unit, King's College Hospital and ¶Imperial College London, Hammersmith Campus, London, UK, **Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Rome, Italy and ††Department of Obstetrics and Gynaecology, Malmö University Hospital, Lund University, Malmö, Sweden*

KEYWORDS: ovarian neoplasms; pattern recognition; risk of malignancy; statistical models; subjective impression; ultrasound

ABSTRACT

Objectives *To determine the effect of an ultrasound training course on the performance of pattern recognition when used by less experienced examiners and to compare the performance of pattern recognition, a logistic regression model and a scoring system to estimate the risk of malignancy between examiners with different levels of experience.*

Methods *Using ultrasound images of selected adnexal masses, two trainees classified the masses as benign or malignant by using pattern recognition both before and after they had attended a theoretical gynecological ultrasound course. They also classified the masses by using a logistic regression model and a scoring system, but only after they had attended the course. The performance of these three methods when they were used by the trainees was then compared with that when they were used by experts.*

Results *One hundred and sixty-five adnexal masses were included, of which 42% were malignant (21% invasive tumors and 21% borderline tumors). The area under the receiver–operating characteristics curve of pattern recognition when used by the trainees was similar before and after they had attended the course. Training decreased sensitivity (84% vs. 70% for Trainee 1, $P = 0.004$; 70% vs. 61% for Trainee 2, $P = 0.058$) and increased specificity (77% vs. 92% for Trainee 1, $P = 0.001$; 89% vs. 95% for Trainee 2, $P = 0.058$). The performance of pattern recognition was poorer in the hands of the trainees than in the hands of the experts. The sensitivities of the logistic regression model were 70% and 54% for the trainees vs. 83% for an expert ($P = 0.020$ and < 0.001, respectively) and the specificities were 84% and 94% vs. 89% ($P = 0.25$ and 0.59, respectively). The sensitivities of the scoring system were 59% and 54% for the trainees vs. 75% for the expert ($P = 0.002$ and < 0.001, respectively), and the specificities were 90% and 93% vs. 85% ($P = 0.103$ and 0.008, respectively).*

Conclusion *Theoretical ultrasound teaching did not seem to improve the performance of pattern recognition in the hands of trainees. A logistic regression model and a scoring system to classify adnexal masses as benign or malignant perform less well when they were used by inexperienced examiners than when used by an expert. Before using a model or a scoring system, experience and/or proper training are likely to be of paramount importance if diagnostic performance is to be optimized.*

INTRODUCTION

Predicting whether an adnexal mass is benign or malignant is pivotal in determining its management (expectant management, laparoscopic surgery, or referral for surgery

Correspondence to: Dr C. Van Holsbeke, Department of Obstetrics and Gynaecology, Z.O.L. Genk, Belgium, Schiepse Bos 6, B-3600 Genk, Belgium (e-mail: caroline.van.holsbeke@skynet.be)

Accepted: 19 February 2009

in a gynecological oncology center). Subjective evaluation of gray-scale and color Doppler ultrasound images of adnexal masses, i.e., pattern recognition, by expert sonologists is an excellent method for discriminating between benign and malignant adnexal masses[1–5]. However, individuals with little or moderate ultrasound experience can discriminate less well between benign and malignant adnexal masses when they use pattern recognition than can experienced ultrasound examiners[6]. A study by Yazbek *et al.* showed that gynecologists with a special interest in sonography influenced decision-making such that there were fewer staging laparotomies and a shorter duration of hospitalization in comparison with level II ultrasound examiners[7]. Other methods than pattern recognition, e.g., the use of logistic regression models to calculate the risk of malignancy in adnexal masses[8,9] or the use of a scoring system to classify adnexal masses as benign or malignant[10–13] might be useful for less experienced ultrasound examiners when they are faced with the task of classifying adnexal masses as benign or malignant. Models and scores may perform well in the hands of experienced ultrasound examiners[8–14], but to the best of our knowledge the diagnostic performance of logistic regression models or scores when used by health professionals with limited ultrasound experience has not been determined.

The aim of this study was to determine the effect of an ultrasound training course on the performance of pattern recognition when used by less experienced examiners and to compare the performance of pattern recognition, a logistic regression model and a scoring system to estimate the risk of malignancy in adnexal masses between examiners with different levels of experience.

METHODS

The study was conducted within the framework of the multicenter International Ovarian Tumor Analysis (IOTA) collaboration[8,12,13,15,16], and the IOTA study was approved by the local ethical committees.

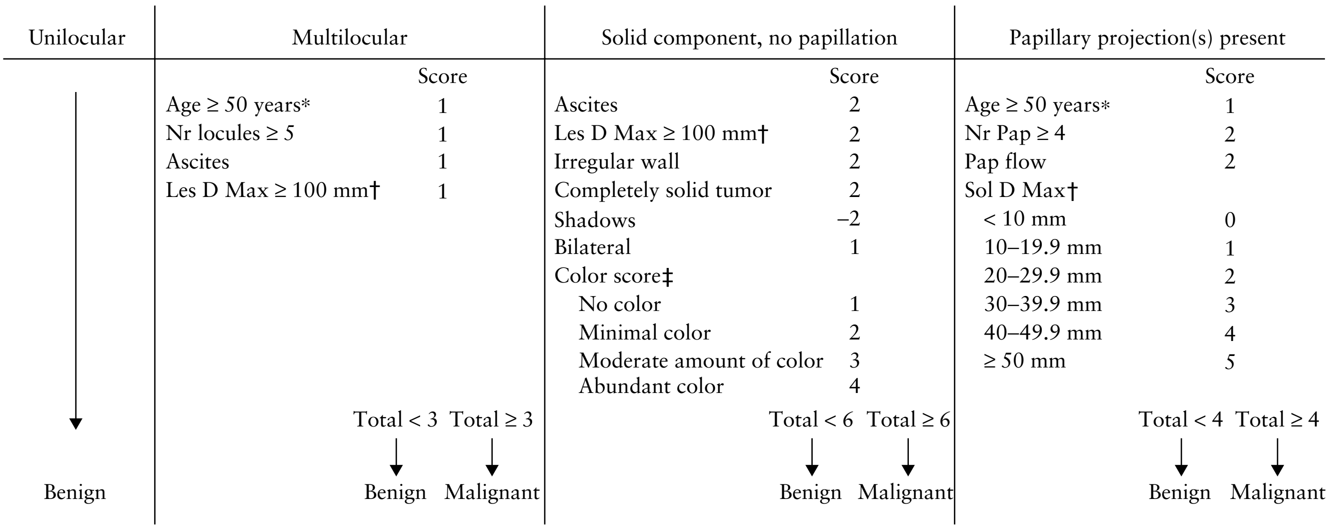
Two trainees in obstetrics and gynecology (T.M., L.L.) attended a theoretical course on gynecological ultrasonography in which they were instructed on how to assess and report on the ultrasound features of an adnexal mass using the terms and definitions published by the IOTA group[16]. To improve their ability to discriminate between benign and malignant adnexal masses using pattern recognition, the ultrasound characteristics of most types of adnexal mass were demonstrated using a large number of ultrasound images. In particular, the ultrasound features of invasive malignant tumors, histological subtypes of borderline tumors[17,18], endometrioma, dermoid cyst, fibroma and hydrosalpinx were described. The benefit of using scoring systems or mathematical models to estimate the risk of malignancy in adnexal masses was discussed, and the main IOTA logistic regression model[8] and an IOTA scoring system[12,13] were discussed in detail. Briefly, the IOTA logistic regression model and scoring system were developed using a database of 1066 patients with an adnexal mass. The data in the database had been prospectively collected within the framework of the IOTA multicenter study phase 1, including information on more than 40 demographic and ultrasound variables. The logistic regression model included the 12 variables shown in Table 1[8]. For the scoring system the masses were categorized into four subgroups based on their ultrasound appearance: (1) unilocular cyst, (2) multilocular cyst, (3) mass with a solid component but no papillary projections, and (4) mass with one or more papillary projections, a papillary projection being defined as a solid structure protruding from the cyst wall and measuring ≥ 3 mm in height[16]. For each of the four subgroups a scoring system is used to classify the tumor as benign or malignant (Figure 1). More information on the logistic regression model and a modified version of the scoring system that we used for this study can be derived from the literature[8,12,13].

The diagnostic performance of pattern recognition, the main IOTA logistic regression model[8] and the IOTA scoring system[12] was tested on a prospectively collected series of electronically saved gray-scale and color Doppler ultrasound images of 165 adnexal masses. For the purpose of this study the images were anonymized. The images came from 166 non-consecutive patients who had been examined preoperatively in the Early Pregnancy and Gynaecology Assessment Unit of King's College Hospital, London, by an expert sonologist between January 2004 and June 2006. One patient was excluded because her images did not contain information on all the ultrasound variables to be used in the logistic regression model and scoring system. Tumors considered to have obviously benign or malignant ultrasound morphology ('easy tumors') were not included in the image collection, the aim being to create a database of tumors that contained a high proportion of difficult-to-classify lesions, because differences in diagnostic performance of pattern recognition between individuals with varying levels of ultrasound experience might be more easy to detect in a population containing a high proportion of difficult tumors[19].

Table 1 Variables in the main IOTA logistic regression model[8]

Age*
Personal history of ovarian cancer*
Largest diameter of lesion†
Largest diameter of largest solid component†
Presence of ascites
Presence of flow in papillary projection
Irregular internal cyst walls
Presence of a purely solid tumor
Color score‡
Presence of acoustic shadows
Current hormonal therapy*
Pain during examination*

*Information on these variables was provided to all examiners. †Measurements that were available in the written report of the real-time ultrasound examiner were used if the images did not provide information on size. ‡If there were no color Doppler images available, the color score assigned by the real-time ultrasound examiner was used.



Unilocular

Benign

Multilocular

	Score
Age ≥ 50 years*	1
Nr locules ≥ 5	1
Ascites	1
Les D Max ≥ 100 mm†	1

Total < 3 → Benign; Total ≥ 3 → Malignant

Solid component, no papillation

	Score
Ascites	2
Les D Max ≥ 100 mm†	2
Irregular wall	2
Completely solid tumor	2
Shadows	−2
Bilateral	1
Color score‡	
No color	1
Minimal color	2
Moderate amount of color	3
Abundant color	4

Total < 6 → Benign; Total ≥ 6 → Malignant

Papillary projection(s) present

	Score
Age ≥ 50 years*	1
Nr Pap ≥ 4	2
Pap flow	2
Sol D Max†	
< 10 mm	0
10–19.9 mm	1
20–29.9 mm	2
30–39.9 mm	3
40–49.9 mm	4
≥ 50 mm	5

Total < 4 → Benign; Total ≥ 4 → Malignant

Figure 1 IOTA subgroup scoring system[12]. *Information on this variable was provided to all examiners. †Measurements that were available in the written report of the expert who had performed the real-time ultrasound examination were used if the images did not show information on size. ‡If there were no color Doppler images available, the color score assigned by the real-time ultrasound examiner was used. Ascites, fluid outside the pouch of Douglas; Color score, color content of the tumor scan at power Doppler examination (no color, minimal color, moderate amount of color, abundant color); Irregular wall, presence of irregular internal walls in the lesion; Les D Max, largest diameter of the lesion; Nr locules, number of locules (0, 1, 2, 3, 4, 5 to 10, or >10); Nr Pap, number of separate papillary projections (1, 2, 3, or >3); Pap flow, color Doppler signals detected in at least one papillary projection; Shadows, presence of acoustic shadows; Sol D Max, largest diameter of the largest solid component.

The images had all been taken during scans performed by an expert sonologist (D.J., one of the IOTA collaborators), who used the IOTA terms and definitions to describe his findings. The images had been taken to demonstrate the most characteristic ultrasound features of the adnexal masses. Because the ability of the examiners to use the model, scoring system and pattern recognition was tested on saved images, we could not test the ability of the examiners to take accurate measurements. If measurement results were not shown on the images, the measurements taken by the expert who had carried out the ultrasound examination were used. Color Doppler images were not available for all the masses, but in all cases the written ultrasound report contained information on the color score. If color Doppler images were not available, the color score given by the expert who carried out the ultrasound examination was used. 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[16]. The histopathological diagnosis of the mass following surgery was the gold standard. The masses were classified using the World Health Organization guidelines for histology[20].

The ultrasound images of the masses were independently evaluated by six reviewing examiners: four ultrasound experts (L.V., A.T., D.T. and C.V.H.) and two trainees in obstetrics and gynecology (T.M. and L.L.). The experts were senior clinicians from different tertiary referral gynecological ultrasound units who had performed at least 5000 ultrasound scans each. The trainees had received at least 6 months' training in gynecological ultrasound in the ultrasound department of one of the experts (D.T.) and had performed over 700 scans each. All six image reviewers received the following clinical information: indication for the scan, symptoms, whether there was a palpable mass present and information on whether there was a personal or family history of breast or ovarian cancer. Three of the ultrasound experts (L.V., A.T. and D.T.) and both trainees evaluated the ultrasound images using pattern recognition, the trainees performing two evaluations, i.e., one before and another one within 1 month after having attended the dedicated ultrasound course described above. Both the experts and the trainees were asked to answer the following questions: (1) 'Based on your subjective impression, do you think the mass is benign or malignant?' (2) 'With which level of confidence do you suggest your diagnosis (certainly benign, probably benign, uncertain, probably malignant, or certainly malignant)?'. The category 'uncertain' reflects the fact that the first question ('Do you think the mass is benign or malignant?') was very difficult to answer. The diagnostic performance of the trainees before and after attending the ultrasound course was compared with that of the 'consensus opinion' of three of the experts (D.T., A.T. and L.V.), the 'consensus opinion' being defined as the diagnosis assigned by at least two of the three experts. In addition, within 1 month after attending the ultrasound course, the two trainees and a fourth ultrasound expert (C.V.H.) independently evaluated the ultrasound images to obtain information on the ultrasound variables used in the main IOTA logistic regression model (Table 1) and in the IOTA score (Figure 1). If the reviewer thought that there was a solid component and the sonologist who had performed the ultrasound examination described the same solid component in his report (as judged from the size of the solid component described in the ultrasound report and the measurements shown on the ultrasound images), then the measurements taken by the original ultrasound examiner were used in the model/score. If the sonologist who performed the ultrasound examination did not describe a solid component in his ultrasound report, then the size of any solid area pointed out by the reviewer on the ultrasound image was estimated from the scale on the ultrasound image by the expert reviewer (C.V.H.), and this estimate was used in the logistic regression model/score. On a list showing the 12 variables used in the logistic regression model (Table 1) the image reviewers noted whether the respective variables were present or not, and the variables noted on the list were used to calculate the risk of malignancy. The risk cut-off (0.1) to indicate malignancy suggested in the publication describing the model was used when classifying a mass as benign or malignant. For the subset scoring system (a modification of which has been published[13]), the three image reviewers used the flowchart shown in Figure 1 to classify the tumor as benign or malignant. The results of each reviewer were compared with the histology of the surgically removed mass. The cases that were correctly classified as malignant by the expert but incorrectly classified as benign by the trainees and those that were correctly classified as benign by the trainees but incorrectly classified as malignant by the expert were scrutinized with the aim of determining which ultrasound variables were interpreted differently by the trainees and the expert reviewer. In this analysis the interpretation of the variables by the expert reviewer was used as the gold standard.

Statistical analysis

Statistical analysis was performed using SAS Version 9.1 for Windows (SAS Institute, Inc., Cary, NC, USA). The sensitivity, specificity and accuracy with regard to malignancy of the logistic regression model, the score and pattern recognition were calculated for the expert sonologists and the trainees. The statistical significance of differences in sensitivity, specificity and accuracy was determined using McNemar's test, which was also used to determine the statistical significance of the differences in sensitivity, specificity and accuracy of pattern recognition before and after theoretical ultrasound education. The area under the receiver–operating characteristics curve (AUC) was calculated for the logistic regression model, the scoring system and pattern recognition, six levels of diagnostic confidence being used to calculate the AUC of pattern recognition (certainly benign, probably benign, uncertain but nevertheless first classified as most likely to be benign, uncertain but nevertheless first classified as most likely to be malignant, probably malignant, and certainly malignant). The statistical significance of differences in AUC was determined using the method of DeLong *et al.*[21]. Two-tailed $P < 0.05$ was considered to indicate a statistically significant difference.

RESULTS

Of the 165 masses included, 69 (42%) were malignant (21% invasive tumors and 21% borderline tumors). Table 2 presents the histopathological diagnoses. The test performances of pattern recognition in the hands of the experts (consensus opinion) and in the hands of the trainees before and after they had undergone the theoretical ultrasound course are shown in Table 3. The AUC for pattern recognition when used by the trainees was similar before and after they had attended the course: it increased slightly for one trainee and decreased slightly for the other, but the differences were not statistically significant (Figure 2). After the course, the sensitivity decreased and the specificity increased. Both before and after the ultrasound course, the performance of pattern recognition was slightly poorer in the hands of the trainees than in the hands of the experts. For both trainees, after the course the sensitivity was statistically significantly

Table 2 Histopathological diagnoses of the masses included

Diagnosis	n *(%)*
Benign	96 (58.2)
Dermoid	35 (21.2)
Cystadenoma/fibroma	35 (21.2)
Endometrioma	16 (9.7)
Fibroma	6 (3.6)
Simple cyst/functional cyst	2 (1.2)
Abscess	1 (0.6)
Rare benign tumor	1 (0.6)
Malignant	69 (41.8)
Mucinous borderline	16 (9.7)
Serous borderline	18 (10.9)
Primary invasive	24 (14.5)
Rare malignant tumor	11 (6.7)

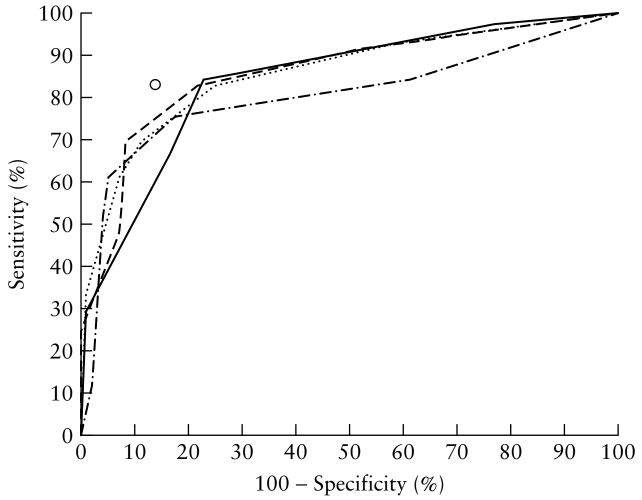


Figure 2 Receiver–operating characteristics curves for pattern recognition when used by two trainees in obstetrics and gynecology before and after they had attended a theoretical ultrasound course. ——, Trainee 1 before course; – – –, Trainee 1 after course; ...., Trainee 2 before course; –·–·–, Trainee 2 after course; ○, consensus opinion.

Table 3 Sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR−) and accuracy with regard to malignancy of subjective evaluation of static ultrasound images when used by expert sonologists and by two trainees in obstetrics and gynecology before and after they had attended a theoretical ultrasound course

Parameter	*Trainee 1*			*Trainee 2*			*Consensus opinion of three experts*‡
	Before course	*After course*	P†	*Before course*	*After course*	P†	
Sensitivity (% (*n*))	84 (58/69), *P** = 0.763	70 (48/69), *P** = 0.020	0.004	70 (48/69), *P** = 0.0201	61 (42/69), *P** < 0.001	0.058	83 (57/69)
Specificity (% (*n*))	77 (74/96), *P** = 0.039	92 (88/96), *P** = 0.059	0.001	89 (85/96), *P** = 0.564	95 (91/96), *P** = 0.011	0.058	86 (83/96)
Accuracy (% (*n*))	80 (132/165), *P** = 0.144	82 (136/165), *P** = 0.343	0.465	81 (133/165), *P** = 0.178	81 (133/165), *P** = 0.194	1	85 (140/165)
LR+	3.65	8.75		6.36	12.2		6.10
LR−	0.21	0.33		0.34	0.41		0.20
AUC	0.840	0.857	0.596	0.854	0.802	0.068	

*Statistical significance of the difference between the trainee and the consensus opinion of the experts (McNemar's test). †Statistical significance of the difference between the results before and after the ultrasound course (McNemar's test for sensitivity, specificity and accuracy and the method described by DeLong *et al.*[21] for AUC). ‡Consensus opinion was defined as the diagnosis suggested by at least two of three expert sonologists. AUC, area under the receiver–operating characteristics curve calculated using six levels of diagnostic confidence.

lower than that of the consensus opinion of the experts, while the specificity was higher, the difference in specificity being statistically significant for one of the trainees.

Tables 4 and 5 present the performance of the main IOTA logistic regression model and the IOTA scoring system when they were used by the two trainees and one expert sonologist. The receiver–operating characteristics curves show that both the model and the score manifested better diagnostic performance when they were used by the expert than by the trainees (Figures 3 and 4). For both the model and the scoring system, the trainees had a lower sensitivity and, in most cases, higher specificity than the expert. To find out which ultrasound features were difficult for the trainees to interpret, the cases where

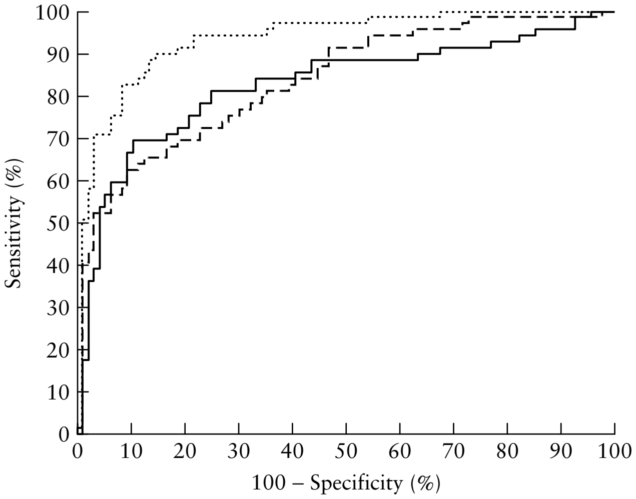


Figure 3 Receiver–operating characteristics curves for the main IOTA logistic regression model[8] for calculating the risk of malignancy in adnexal masses when used by three sonologists – two trainees in obstetrics and gynecology and one ultrasound expert. ——, Trainee 1; - - -, Trainee 2; ...., expert.

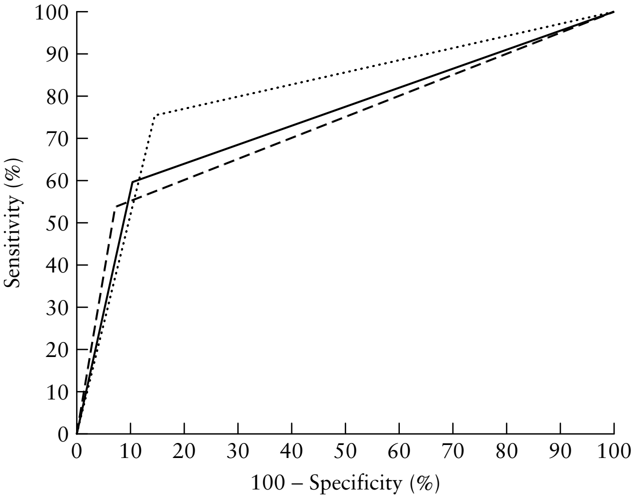


Figure 4 Receiver–operating characteristics curves for the IOTA scoring system[12] for classifying adnexal masses as benign or malignant when used by three sonologists – two trainees in obstetrics and gynecology and one ultrasound expert. ——, Trainee 1; - - -, Trainee 2; ...., expert.

Table 4 Sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR−) and accuracy of the IOTA logistic regression model[8] when used by two trainees in obstetrics and gynecology and an expert ultrasound examiner evaluating static ultrasound images

Sonologist	*AUC*	P*	*Sensitivity (% (*n*))*	P*	*Specificity (% (*n*))*	P*	*LR+*	*LR−*	*Accuracy (% (*n*))*	P*
Trainee 1	0.827	< 0.001	70 (48/69)	0.020	84 (81/96)	0.248	4.45	0.36	78 (129/165)	0.012
Trainee 2	0.835	< 0.001	54 (37/69)	< 0.001	94 (90/96)	0.058	8.58	0.49	77 (127/165)	0.005
Expert	0.934		83 (57/69)		89 (85/96)		7.21	0.20	86 (142/165)	

*Statistical significance of differences between the trainees and the expert (McNemar's test used for differences in sensitivity, specificity and accuracy; method of DeLong *et al.*[21] used for differences in AUC). AUC, area under the receiver–operating characteristics curve.

Table 5 Sensitivity, specificity, positive likelihood ratio (LR+), negative likelihood ratio (LR−) and accuracy of the IOTA scoring system[12] when used by two trainees in obstetrics and gynecology and an expert ultrasound examiner evaluating static ultrasound images

Sonologist	*AUC*	P*	*Sensitivity (% (*n*))*	P*	*Specificity (% (*n*))*	P*	*LR+*	*LR−*	*Accuracy (% (*n*))*	P*
Trainee 1	0.745	0.033	59 (41/69)	0.002	90 (86/96)	0.103	5.70	0.45	77 (127/165)	0.108
Trainee 2	0.732	0.017	54 (37/69)	< 0.001	93 (89/96)	0.008	7.36	0.50	76 (126/165)	0.103
Expert	0.804		75 (52/69)		85 (82/96)		5.17	0.29	81 (134/165)	

*Statistical significance of differences between the trainees and the expert (McNemar's test used for differences in sensitivity, specificity and accuracy; method of DeLong *et al.*[21] used for differences in AUC). AUC, area under the receiver–operating characteristics curve.

the model/score in the hands of the expert resulted in a correct diagnosis with regard to malignancy (true positive for the expert) but not in the hands of the trainees (false negative for trainees) were scrutinized as well as the cases where the model correctly predicted benignity in the hands of the trainees (true negative) but not in the hands of the expert (false positive for the expert). Discrepancies with regard to the presence of solid components, wall irregularity and acoustic shadowing explained most differences in classification between the trainees and the expert when they used the IOTA logistic regression model. Wall irregularity and acoustic shadowing explained most differences in classification between the trainees and the expert when they used the IOTA score (Tables 6 and 7).

DISCUSSION

To determine the risk of malignancy in an adnexal mass, the best approach currently available is pattern recognition used by an expert sonologist[3,4,6,7]. The level of ultrasound expertise as well as clinical experience of the person carrying out a scan will impact on the quality of an ultrasound examination, and this in turn may influence the management of patients[7]. The main purpose of developing mathematical models or scoring systems to estimate the risk of malignancy in an adnexal mass is to give less experienced ultrasound examiners a tool to achieve the same diagnostic performance as an expert. In studies evaluating the performance of mathematical models, the examiners who performed the ultrasound examinations and validated the models were experts[4,8–15]. In this study we investigated the performance of pattern recognition when used by sonologists with different levels of ultrasound experience, and examined the impact of training on the performance of examiners with limited experience. We also examined the effect of the level of experience of ultrasound examiners on the diagnostic performance of a logistic regression model and a scoring system developed to estimate the risk of malignancy.

Arger *et al.*[22] demonstrated that four 2-hour training sessions of medical students on ultrasound of the aorta and kidney significantly increased their basic knowledge of sonography and improved their scanning skills. A theoretical and practical training course in musculoskeletal ultrasonography also demonstrated a significant post-course improvement in ultrasound skills[23]. Surprisingly, we were unable to demonstrate that attending an ultrasound course improved the performance of pattern recognition when used by trainees in obstetrics and gynecology with some but limited ultrasound experience. In fact, attending the course resulted in a significant decrease in the sensitivity of pattern recognition. Moreover, despite training, the performance of the logistic regression model and the scoring system was poorer in the hands of the trainees than in the hands of an expert. Indeed, in the hands of the trainees, both methods were associated with too low a sensitivity for the methods to be clinically useful. This seemed to be explained by the trainees failing to recognize ultrasound features typical of malignancy. Perhaps the course focused too little on explaining the ultrasound features to be included in the model/score.

It is a limitation of our study that the performance of pattern recognition, the scoring system and mathematical model was evaluated using static ultrasound images. However, it would have been difficult to submit patients to examination by three or more examiners (in this case six examiners), as would have been necessary to evaluate the ability of all the examiners not only to interpret ultrasound images but also to create them. The

Table 6 Discrepancies explaining why trainees had more false-negative cases than the expert and the expert had more false-positive cases than the trainees when they tested the main IOTA logistic regression model[8]

	False negatives by trainees/ true positives by expert		*False positives by expert/ true negatives by trainees*	
How ultrasound variables were misinterpreted by trainees	*Trainee 1 (n = 10)*	*Trainee 2 (n = 18)*	*Trainee 1 (n = 4)*	*Trainee 2 (n = 6)*
Forgot to include personal history of ovarian cancer	3	4	—	—
Did not recognize solid component	7	16	1	3
Did not recognize that mass was purely solid	—	1	1	1
Did not recognize irregularity of cyst wall	9	17	1	4
Did not see flow inside papillary projection	1	3	—	—
Misinterpretation of acoustic shadowing	3	10	2	1
Did not recognize presence of ascites	1	—	—	—

Table 7 Discrepancies explaining why trainees had more false-negative cases than the expert and the expert had more false-positive cases than the trainees when they tested the IOTA subset scoring system[12]

	False negatives by trainees/ true positives by expert		*False positives by expert/ true negatives by trainees*	
How ultrasound variables were misinterpreted by trainees	*Trainee 1 (n = 12)*	*Trainee 2 (n = 16)*	*Trainee 1 (n = 5)*	*Trainee 2 (n = 7)*
Assignment of wrong locularity	11	11	3	6
Classified as unilocular but it was multilocular	3	—	2	4
Classified as unilocular but it was multilocular-solid	2	5	1	1
Classified as unilocular but it was unilocular-solid	—	—	—	1
Classified as multilocular-solid but did not recognize papillary projections	6	6	—	—
Did not recognize (correct size of) solid component	2	8	2	2
Did not recognize that mass was purely solid	—	—	—	—
Did not recognize irregularity of cyst wall	5	6	—	—
Did not see flow inside papillary projection	1	—	—	—
Incorrect interpretation of acoustic shadowing	4	3	—	—
Did not recognize that number of locules was ≥ 5	—	1	2	3
Did not recognize that there were ≥ 4 papillary projections	—	1	—	—
Ignored that lesion was >100 mm	—	—	—	1

performance of the methods might have been poorer if the trainees had needed to perform the ultrasound examinations themselves, because experience is needed to produce representative images; or it might have been better, because a real-time examination is likely to be more informative than still images[24]. The use of still images also made it impossible to evaluate the ability of the reviewers to measure an adnexal mass and to assign a color score (color Doppler images often not being available).

To sum up, our results show that it is difficult for less experienced examiners to replicate the performance of expert sonologists. Our data suggest that – at least when based on the interpretation of static images – not only pattern recognition but also logistic regression models and scoring systems to estimate the risk of malignancy in adnexal masses do not perform well in the hands of examiners with limited ultrasound experience, in all likelihood because they fail to recognize characteristic ultrasound features. It is obvious that we need to develop ways of teaching less experienced operators how to interpret ultrasound images. Moreover, each course aiming at improving the ultrasound skills of the participants probably needs to include hands-on training as well. Before using a model or a scoring system, proper training is likely to be of paramount importance if diagnostic performance is to be optimized.

ACKNOWLEDGMENTS

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); Funds administered by Malmö University Hospital.

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. Use of morphology to characterize and manage common adnexal masses. *Best Pract Res Clin Obstet Gynaecol* 2004; **18**: 71–89.
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. 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.
5. Van Calster B, Timmerman D, Bourne T, Testa A, Van Holsbeke C, Domali E, Jurkovic D, Neven P, Van Huffel S, Valentin L. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125. *J Natl Cancer Inst* 2007; **99**: 1706–1714.
6. Timmerman D, Schwarzler P, Collins WP, Claerhout F, Coenen M, Amant F, Vergote I, Bourne TH. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet Gynecol* 1999; **13**: 11–16.
7. 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.
8. 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.
9. Tailor A, Jurkovic D, Bourne T, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. *Ultrasound Obstet Gynecol* 1997; **10**: 41–47.
10. Lerner JP, Timor-Tritsch IE, Federman A, Abramovich G. Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring system. *Am J Obstet Gynecol* 1994; **170**: 81–85.
11. Sassone AM, Timor-Tritsch I, Artner A, Westhoff C, Warren W. Transvaginal sonographic characterization of ovarian disease: evaluation of a new scoring system to predict ovarian malignancy. *Obstet Gynecol* 1991; **78**: 70–76.
12. Ameye L, Valentin L, Testa AC, Van Holsbeke C, Domali E, Van Huffel S, Vergote I, Bourne T, Timmerman D. 'Scoring system to differentiate benign from malignant masses in specific subgroups of adnexal tumors', Internal Report 08–114, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2008.
13. Ameye L, Valentin L, Testa AC, Van Holsbeke C, Domali E, Van Huffel S, Vergote I, Bourne T, Timmerman D. A scoring system to differentiate malignant from benign masses in specific ultrasound-based subgroups of adnexal tumors. *Ultrasound Obstet Gynecol* 2009; **33**: 92–101.
14. DePriest P, Shenson D, Fried A, Hunter J, Andrews S, Gallion H, Pavlik E, Kryscio R, van Nagell J. A morphology index based on sonographic findings in ovarian cancer. *Gynecol Oncol* 1993; **51**: 7–11.
15. 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 (IOTA) study. *Clin Cancer Res* 2009; **15**: 684–691.
16. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I; International Ovarian Tumor Analysis (IOTA) Group. Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound Obstet Gynecol* 2000; **16**: 500–505.
17. Fruscella E, Testa AC, Ferrandina G, De Smet F, Van Holsbeke C, Scambia G, Zannoni GF, Ludovisi M, Achten R, Amant F, Vergote I, Timmerman D. Ultrasound features of different histopathological subtypes of borderline ovarian tumors. *Ultrasound Obstet Gynecol* 2005; **26**: 644–650.
18. Valentin L, Ameye L, Testa A, Lécuru F, Bernard JP, Paladini D, Van Huffel S, Timmerman D. Ultrasound characteristics of different types of adnexal malignancies. *Gynecol Oncol* 2006; **102**: 41–48.
19. 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.
20. Senoy SF, Scully RE, Sobin LH. *The World Health Organization international histological classification of ovarian tumours*. World Health Organization: Geneva, 1973.
21. 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.
22. Arger P, Schultz S, Sehgal C, Cary T, Aronchick J. Teaching medical students diagnostic sonography. *J Ultrasound Med* 2005; **24**: 1365–1369.
23. Brown A, Wakefield R, Karim Z, Roberts T, O'Connor P, Emery P. Evidence of effective and efficient teaching and learning strategies in the education of rheumatologist ultrasonographers: evaluation from the 3rd BSR musculoskeletal ultrasonography course. *Rheumatology* 2005; **44**: 1068–1069.
24. 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.