Ultrasound Obstet Gynecol 2009; 33: 92–101
Published online 17 December 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/uog.6273

A scoring system to differentiate malignant from benign masses in specific ultrasound-based subgroups of adnexal tumors

L. AMEYE*, L. VALENTIN†, A. C. TESTA‡, C. VAN HOLSBEKE§, E. DOMALI§, S. VAN HUFFEL*, I. VERGOTE§, T. BOURNE§ and D. TIMMERMAN§

**Department of Electrical Engineering (ESAT-SCD) and §Department of Obstetrics and Gynecology, University Hospitals, Katholieke Universiteit Leuven, Leuven, Belgium, †Department of Obstetrics and Gynecology, Malmö University Hospital, Lund University, Malmö, Sweden and ‡Instituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Rome, Italy*

KEYWORDS: color Doppler imaging; ovarian neoplasms; ultrasonography

ABSTRACT

Objective To investigate if the prediction of malignant adnexal masses can be improved by considering different ultrasound-based subgroups of tumors and constructing a scoring system for each subgroup instead of using a risk estimation model applicable to all tumors.

Methods We used a multicenter database of 1573 patients with at least one persistent adnexal mass. The masses were categorized into four subgroups based on their ultrasound appearance: (1) unilocular cyst; (2) multilocular cyst; (3) presence of a solid component but no papillation; and (4) presence of papillation. For each of the four subgroups a scoring system to predict malignancy was developed in a development set consisting of 754 patients in total (respective numbers of patients: (1) 228; (2) 143; (3) 183; and (4) 200). The subgroup scoring system was then tested in 312 patients and prospectively validated in 507 patients. The sensitivity and specificity, with regard to the prediction of malignancy, of the scoring system were compared with that of the subjective evaluation of ultrasound images by an experienced examiner (pattern recognition) and with that of a published logistic regression (LR) model for the calculation of risk of malignancy in adnexal masses. The gold standard was the pathological classification of the mass as benign or malignant (borderline, primary invasive, or metastatic).

Results In the prospective validation set, the sensitivity of pattern recognition, the LR model and the subgroup scoring system was 90% (129/143), 95% (136/143) and 88% (126/143), respectively, and the specificity was 93% (338/364), 74% (270/364) and 90% (329/364), respectively.

Conclusions In the hands of experienced ultrasound examiners, the subgroup scoring system for diagnosing malignancy has a performance that is similar to that of pattern recognition, the latter method being the best diagnostic method currently available. The scoring system is less sensitive but more specific than the LR model.
Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

INTRODUCTION

Malignant ovarian tumors are diagnosed at an advanced stage in 75% of cases and are associated with the highest mortality figures of all gynecological cancers[1]. It may be difficult to determine preoperatively the nature (benign or malignant) of adnexal tumors. However, an accurate diagnosis is essential to provide optimal treatment as the rupture of a Stage I ovarian cancer during surgery may worsen the prognosis[2]. Good preoperative discrimination between benign and malignant ovarian tumors results in more women being appropriately referred for gynecologic oncology care and more women with benign conditions undergoing conservative surgical treatment[3].

In 1989 Granberg *et al.* reported that the gross morphology of adnexal tumors could be used to predict the likelihood of malignancy[4]. They also found that ultrasound images of tumors predicted their gross morphology, and they suggested that therefore it should be possible to estimate the risk of malignancy on the basis of ultrasound morphology[4]. Indeed, this has proven to be true, and subjective evaluation of ultrasound findings (pattern recognition) by an experienced ultrasound examiner is an excellent method for discriminating between benign and malignant adnexal tumors[5–7]. In 1990 Jacobs *et al.* incorporated ultrasound findings, the patient's menopausal status and serum CA-125 level into a risk-of-malignancy index[8]. Later, scoring systems[9], logistic regression (LR) models[10,11] and neural

Correspondence to: Prof. D. Timmerman, Department of Obstetrics and Gynecology, University Hospitals, KU Leuven, Herestraat 49, B-3000 Leuven, Belgium (e-mail: dirk.timmerman@uz.kuleuven.ac.be)

Accepted: 10 October 2008

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

networks[12,13] were developed to discriminate between benign and malignant tumors. Usually these models were created in single centers using small sample sizes, and there was no consistency in the interpretation of ultrasound-derived variables[14]. When the scoring systems and mathematical models were validated prospectively in different centers they did not perform very well[15–17].

In 1999 a prospective, European multicenter study including nine centers from five countries (Belgium, Sweden, Italy, France and the UK) was set up, the so-called International Ovarian Tumor Analysis (IOTA) study. Its aim was to minimize the limitations of previous work by prospectively collecting the history and ultrasound findings of more than 1000 patients with a persistent mass following a standardized protocol (1999–2002)[18]. Information on more than 50 explanatory variables was collected. An LR model with 12 variables was created to calculate the risk of malignancy in an adnexal mass. It had a sensitivity of 93% and a specificity of 76% in an independent test set[19]. Below, this model is called the 'overall LR model', the argument for using this term being that the model was developed to be applicable to all ovarian masses.

When we constructed the overall LR model, missing structural values in the data were imputed by zero, e.g. if there was no papillation, the height of the largest papillary structure was imputed with a zero[19]. The overall LR model contains two variables for which imputation was needed: the maximal diameter of the largest solid component and the presence of blood flow within a papillation; if there was no solid component or no papillation, a zero was imputed. Thus, different subgroups of tumors can be considered on the basis of there being structural missing values in the data, e.g. a subgroup of masses with papillation and another subgroup of masses with solid components but no papillation. Using different prediction models for subgroups of masses with different ultrasound morphology instead of using one single model for all tumors might improve discrimination between benign and malignant tumors.

The aim of the present study was to investigate if a scoring system developed for subgroups of tumors would have better diagnostic performance for discriminating between benign and malignant adnexal masses than one LR model applicable to all tumors.

PATIENTS AND METHODS

The patients included in this study were the 1066 patients of the first phase of the IOTA study (1999–2002)[19] as well as 507 patients from a follow-up study (2002–2006) in which three of the nine original IOTA centers participated (Belgium, Sweden and Italy). Patients presenting with at least one adnexal mass, who underwent an ultrasound examination by a principal investigator at one of the participating centers, were eligible for inclusion. Exclusion criteria were pregnancy, inability to tolerate transvaginal sonography, and surgery performed more than 120 days after sonographic assessment. The choice of the 120-day interval between the scan and surgery was arbitrary, but we assume that the shorter the interval between the scan and the surgical removal of the mass, the less likely would there be a change in the ultrasound morphology or histology (e.g. from benign to borderline, or from borderline to malignant) during that interval. When two masses were present, only information from the most complex one was used in the statistical analysis. The IOTA protocol required the ultrasound examiner not only to prospectively collect clinical and ultrasound information in a standardized manner but also to classify each mass as benign or malignant on the basis of pattern recognition[5]. The examination technique and definitions of terms have been described in previous publications[18,19]. Information on more than 50 clinical and ultrasonographic variables was prospectively collected and considered for association with the outcome (benign, malignant)[19], and information on all these variables was available in all cases (Table 1). The centers were also encouraged to measure the level of serum CA-125 in peripheral blood from all patients, but the availability of this biochemical end-point was not an essential requirement for recruitment into the study. The gold standard was the histopathological classification of the mass as benign, borderline, primary invasive or metastatic. In order to check the quality of the original histopathology report, Prof. P. Moerman, who specializes in gynecological histopathology (Katholieke Universiteit Leuven, Belgium) reviewed a random selection of original pathology blocks from all the participating centers.

The overall LR model containing 12 variables was constructed using a development set of 70% of the data of the first phase of the IOTA study ($n = 754$) and tested on a test set containing the remaining 30% of patients ($n = 312$)[19]. The scoring system for specific subgroups of masses was constructed using the same development and test set and using the data of the follow-up study (2002–2006) as a prospective validation set. In order to enhance the prediction of the overall LR model, subgroups were constructed on the basis of there being structural missing values in the data as described in the Introduction. To create the subgroups the first criterion was the presence or absence of a solid component within the mass. In the absence of solid components the masses were further divided on the basis of number of locules into unilocular and multilocular. Masses with a solid component were further subdivided on the basis of the presence or absence of papillation(s) (Figure 1). Thus, four subgroups were created: (1) unilocular cyst; (2) multilocular cyst; (3) tumor with at least one solid component but no papillation; and (4) tumor with papillation. Ultrasound images representative of tumors in the four subgroups are shown in Figure 2.

For each subgroup an LR model was constructed using the development set of the IOTA phase 1 dataset[19]. In each subgroup, for all possible combinations of variables, the LR model and the associated likelihood score Chi-square statistic was calculated, i.e. for all possible LR models containing only one variable, for all possible LR models

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

Table 1 Variables considered for association with the outcome (benign, malignant)

Continuous demographic characteristics
Age
Number of years postmenopause
Parity
Continuous gray-scale ultrasound findings
Maximal diameter of the lesion (mm)
Volume of lesion (mL)
Fluid in the anteroposterior plane of the pouch of Douglas (mm)
Thickness of thickest septum where it appeared to be at its thickest (mm)
Height of largest papillary projection (mm)
Maximal diameter of papillary projection (mm)
Ratio between the volume of the largest papillary projection and the volume of the lesion
Number of separate papillary projections (1, 2, 3 or > 3)
Number of locules (0, 1, 2, 3, 4, 5–10 or > 10)
Maximal diameter of the largest solid component (mm)
Volume of the largest solid component (mL)
Ratio between the volume of the largest solid component and the volume of the lesion
Continuous blood flow indices
Pulsatility index
Resistance index
Peak systolic velocity (cm/s)
Time-averaged maximum velocity (cm/s)
Tumor marker CA-125 (U/mL)
Binary variables
First-degree relatives with ovarian cancer
First-degree relatives with breast cancer
Personal history of ovarian cancer
Personal history of breast cancer
Nulliparous
Hysterectomy
Postmenopausal
Hormone therapy
Postmenopausal bleeding within 1 year before the ultrasound examination
Bilateral lesions
Pain during the ultrasound examination
Ascites
Incomplete septum
Papillation
Flow within at least one of the papillary projections
Irregular papillary projection
Irregular cyst walls
Acoustic shadows
Venous blood flow only
Categorical characteristics
Tumor type (unilocular, unilocular-solid, multilocular, multilocular-solid, solid, not classifiable)
Echogenicity of cyst fluid (anechoic, low level, ground glass, hemorrhagic, mixed echogenicity, no cyst fluid)
Color score (no flow, minimal flow, moderately strong flow, very strong flow according to subjective evaluation of the color content of the tumor scan at color Doppler examination)

containing two variables and so on. The LR models with the highest score Chi-square were retained and in those we selected the model that fitted with clinical experience. So, for each subgroup one LR model was obtained. A scoring system was derived from the maximum likelihood odds ratio (OR) estimates of the LR model. OR estimates larger than 1 were rounded to the nearest integer. ORs between 0 and 1 were transformed into −(1/(OR)) and

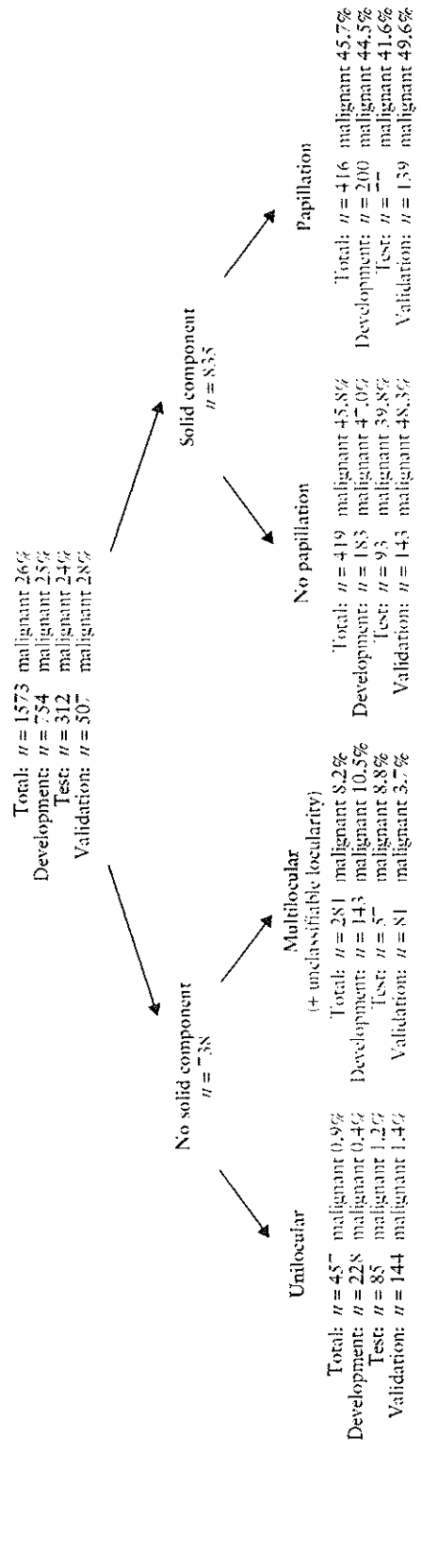


Figure 1 Prevalence of malignancy in each subgroup.

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

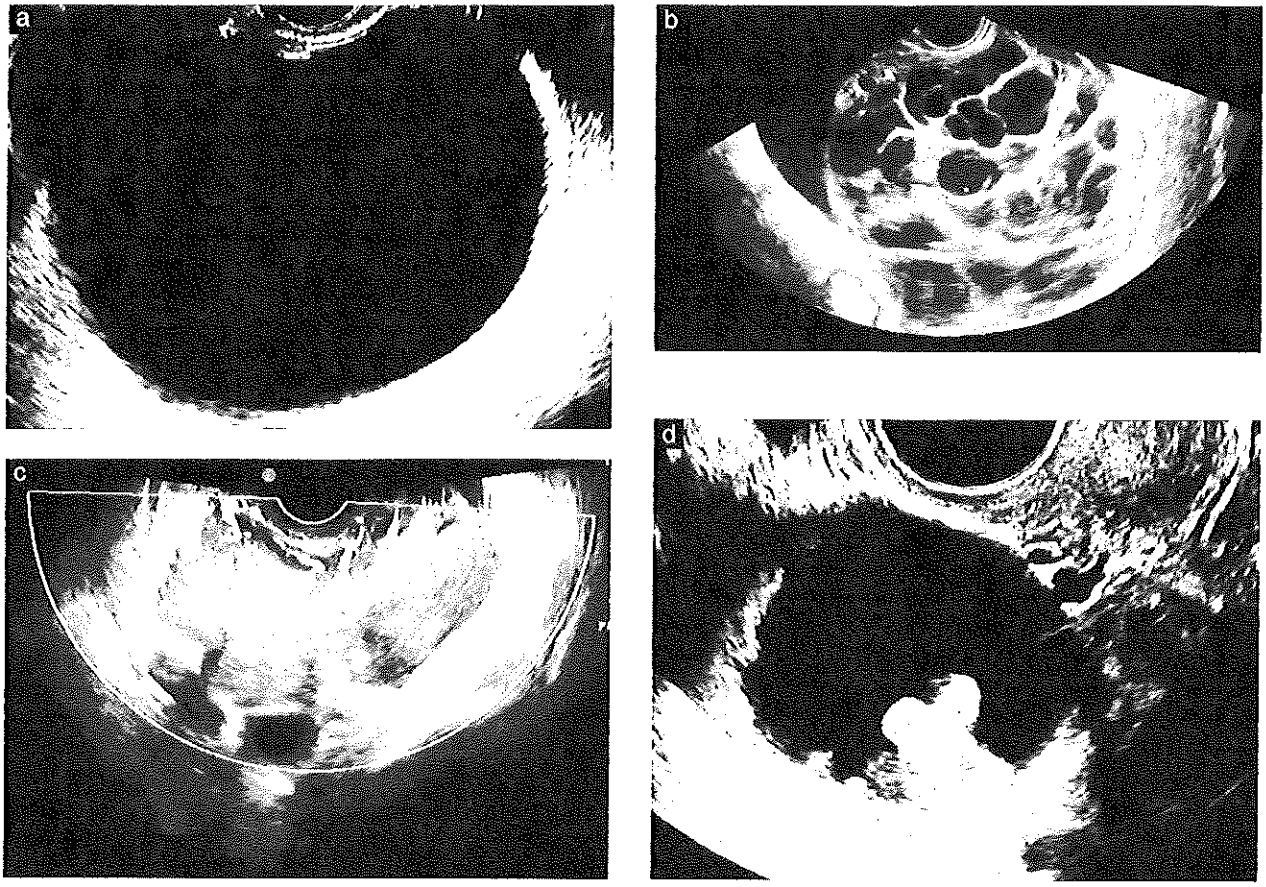


Figure 2 Ultrasound images showing: a unilocular cyst in a 19-year-old patient with a teratoma (a), a multilocular cyst in a 28-year-old patient with a mucinous borderline ovarian tumor (b), a tumor with a solid component but no papillation in a 47-year-old patient with a primary invasive ovarian cancer (c) and a cyst with papillation in a 17-year-old patient with a serous borderline ovarian tumor (d).

then rounded to the nearest integer. The total score was the sum of these individual weights. The cut-offs of our subgroup scoring system to indicate malignancy were chosen so as to result in a clinically reasonable detection rate of malignancy.

The diagnostic performance of pattern recognition, our previously published overall LR model and the subgroup scoring system was expressed in terms of sensitivity, specificity and positive and negative likelihood ratios. When we calculated the sensitivity and specificity of the overall LR model we used the risk cut-off (10%) recommended by Timmerman *et al.*[19]. McNemar's test was used to test the statistical significance of differences in sensitivity and specificity between pattern recognition, the overall LR model and the subgroup scoring system. The statistical significance of differences in demographic background variables between the development set, test set and prospective validation set was determined using the Kruskal–Wallis test for continuous data and the Chi-squared test or Fisher's exact test for discrete data. Two-tailed $P < 0.05$ was considered to indicate a statistically significant difference.

RESULTS

Histological diagnoses are shown in Table 2. The malignancy rate in the total study sample was 26% (409/1573) with 75 (18%) of the 409 malignancies being borderline tumors, 272 (67%) primary invasive tumors, and 62 (15%) metastatic tumors. These rates were similar in the development set, test set and prospective validation set. There were no statistically significant differences in demographic background variables between the three sets (Table 3).

There were 457 unilocular cysts, of which four (0.9%) were malignant (one of 228 in the development set, one of 85 in the test set, and two of 144 in the prospective validation set), 281 multilocular cysts, of which 23 (8%) were malignant, 419 tumors with solid components but no papillation, of which 192 (46%) were malignant and 416 tumors with papillation, of which 190 (46%) were malignant.

The four malignant unilocular cysts were three Stage I borderline tumors and one primary invasive Stage III tumor. These malignancies had a largest diameter of 58 mm, 128 mm, 140 mm and 57 mm, respectively. Because a unilocular cyst was highly predictive of benignity, no scoring system was developed for this

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

Table 2 Histology of the tumors (n = 1573 patients)

Parameter	Development set (N = 754) (n (%))	Test set (N = 312) (n (%))	Prospective validation set (N = 507) (n (%))	Total (N = 1573) (n (%))
Benign pathology				
Endometrioma	147 (19.5)	65 (20.8)	101 (19.9)	313 (19.9)
Dermoid/teratoma	84 (11.1)	37 (11.9)	55 (10.9)	176 (11.2)
Simple cyst*	97 (12.9)	33 (10.6)	59 (11.6)	189 (12.0)
Hydrosalpinx†	27 (3.6)	8 (2.6)	16 (3.2)	51 (3.2)
Peritoneal pseudocyst	3 (0.4)	2 (0.6)	5 (1.0)	10 (0.6)
Abscess	12 (1.6)	2 (0.6)	4 (0.8)	18 (1.1)
Fibroma‡	26 (3.5)	11 (3.5)	29 (5.7)	66 (4.2)
Serous cystadenoma	88 (11.7)	48 (15.4)	48 (9.5)	184 (11.7)
Mucinous cystadenoma	69 (9.2)	25 (8.0)	38 (7.5)	132 (8.4)
Rare benign§	10 (1.3)	6 (1.9)	9 (1.8)	25 (1.6)
Malignant pathology				
Primary invasive	121 (16.0)	48 (15.4)	103 (20.3)	272 (17.3)
Stage I	31	11	24	66
Stage II	9	3	5	17
Stage III	56	17	59	132
Stage IV	9	8	11	28
Rare malignant¶	16	9	4	29
Borderline	40 (5.3)	15 (4.8)	20 (3.9)	75 (4.8)
Stage I	35	15	15	65
Stage II	5	—	—	5
Stage III	—	—	5	5
Metastatic	30 (4.0)	12 (3.9)	20 (3.9)	62 (3.9)

*Including parasalpingeal cyst, inclusion cyst and normal ovary. †Including salpingitis. ‡Including one case of ovarian leiomyoma and twelve cases of uterine leiomyoma or lipoleiomyoma, the latter having been misdiagnosed as an ovarian mass by the ultrasound examiner. §Including struma ovarii, Brenner tumor, Sertoli cell tumor, stromal cell tumor, Schwannoma and lymphangioma. ¶Including granulosa cell tumor, Leydig cell tumor, dysgerminoma, gynandroblastoma, leiomyosarcoma, immature teratoma, malignant mixed Müllerian tumor, small cell cancer, Brenner cancer, carcinosarcoma, choriocarcinoma and yolk sac tumor. N, total number in the set; n, number in the subgroup.

Table 3 Demographic data of the study population (n = 1573)

Parameter	Total (n = 1573)	Development set (n = 754)	Test set (n = 312)	Prospective validation set (n = 507)	P for the differences between the three sets
Age (years)	46 (9–94)	47 (17–94)	45 (17–90)	46 (9–90)	0.88
Parity	1 (0–10)	1 (0–10)	1 (0–8)	1 (0–9)	0.91
Postmenopausal	635 (40)	311 (41)	121 (39)	203 (40)	0.74
Hormonal therapy	332 (21)	176 (23)	59 (19)	97 (19)	0.11
Hysterectomy	116 (7)	58 (8)	20 (6)	38 (8)	0.76
Family history of ovarian cancer*	47 (3)	23 (3)	10 (3)	14 (3)	0.93
Family history of breast cancer†	174 (11)	79 (10)	40 (13)	55 (11)	0.53
Personal history of ovarian cancer	21 (1)	11 (1)	3 (1)	7 (1)	0.92
Personal history of breast cancer	54 (3)	28 (4)	10 (3)	16 (3)	0.89

Values are median (range) or n (%). *Number of women with first-degree relatives with ovarian cancer. †Number of women with first-degree relatives with breast cancer.

subgroup of tumor. Instead, all unilocular cysts were classified as benign. Moreover, there was only one malignant mass in the development set of unilocular tumors, and so it was impossible to develop a risk estimation model for unilocular cysts. The scoring system developed for multilocular cysts contained four variables, while the system for masses with solid components but no papillations and that for masses with papillations both contained eight variables (Figure 3).

Table 4 presents the diagnostic performance of the published overall LR model, the subgroup scoring system and pattern recognition (the subjective assessment by experienced ultrasound examiners). The results are shown for the development set, the test set and the prospective validation set. In the subset of masses with a solid component but no papillations and in that of tumors with papillation, the specificity of the respective subgroup score was significantly higher than that of the overall LR model (specificity

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

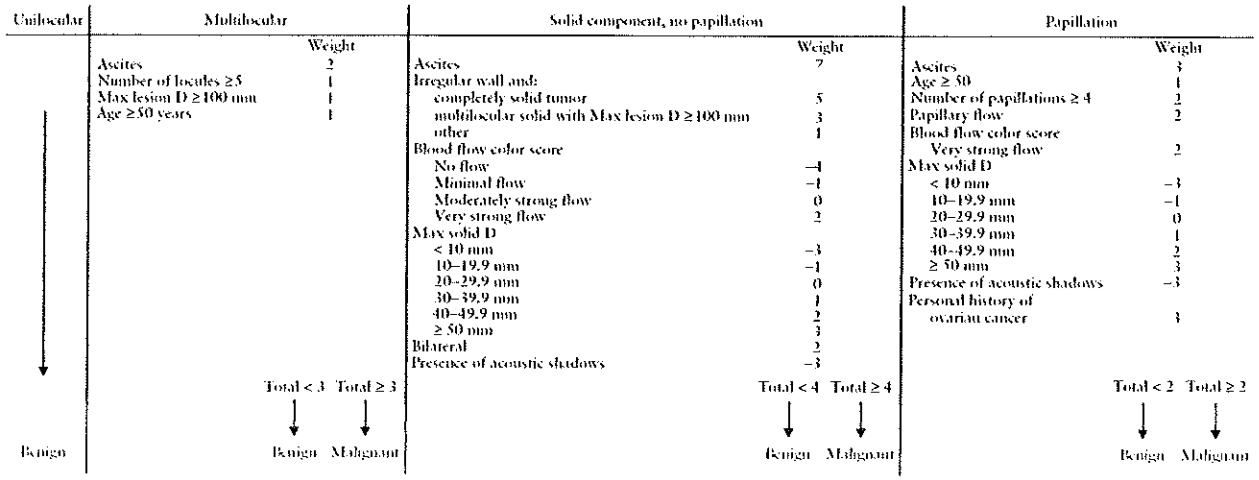


Figure 3 Subgroup scoring system. Ascites, fluid outside the Pouch of Douglas; Irregular wall, presence of irregular internal walls in the lesion; Max lesion D, maximal diameter of the lesion; Max solid D, maximal diameter of the solid component; Papillary flow, color Doppler signals detected in at least one papillary projection.

in the prospective validation set 85% vs. 46% for masses with solid components but no papillation, $P < 0.001$; and 74% vs. 49% for masses with papillation, $P < 0.001$). However, the specificity of the subgroup score was only slightly lower and not statistically significantly different from that of pattern recognition (specificity 85% vs. 88% for masses with solid components but no papillations, $P = 0.59$; and 74% vs. 81% for masses with papillation, $P = 0.33$). For tumors with papillations the sensitivity of the subgroup scoring system was similar to that of the overall LR model and that of pattern recognition (96% vs. 99%, $P = 0.32$; and 96% vs. 97%, $P = 0.56$), but for masses with a solid component but no papillations the overall LR model had significantly higher sensitivity than the subgroup score and pattern recognition (97% vs. 84% and 87%, both $P < 0.01$). The results for the prospective validation set show that the use of the subgroup scoring system would result in sensitivity and specificity similar to that of pattern recognition (sensitivity 88% vs. 90%, $P = 0.44$; specificity 90% vs. 93%, $P = 0.12$) and in a 7% lower sensitivity but 16% higher specificity than the use of the overall LR model (sensitivity 88% vs. 95%, $P = 0.01$; specificity 90% vs. 74%, $P < 0.001$).

Table 5 presents the ability of the overall LR model, the subgroup scoring system and pattern recognition to correctly classify a mass as benign or malignant depending on its specific histology. Pattern recognition, the overall LR model and the subgroup scoring system tended to misclassify abscesses, fibromas and rare benign tumors as malignant, in particular the overall LR model. On the other hand, the overall LR model correctly classified a greater proportion of the borderline tumors as malignant than both pattern recognition and the subgroup scoring system (79% vs. 67%, $P = 0.02$ and 79% vs. 69%, $P = 0.07$, respectively).

Table 6 shows the prevalence of benign and malignant tumors for the possible values of the subgroup scoring system. As expected, owing to the way in which the score was constructed, the higher the score, the greater is the risk of a malignant adnexal mass. Even a low score is associated with a certain risk of malignancy, e.g., in tumors with papillations and an associated total score of 1, still 16% (5/31) were malignant – one borderline stage I and four primary invasive masses (one Stage I, one Stage II and two Stage III).

DISCUSSION

Our data suggest that the use of a scoring system for subgroups of tumors instead of the use of one and the same LR model for all tumors may improve discrimination between benign and malignant adnexal tumors. Indeed, our scoring system matches the diagnostic performance of pattern recognition. In agreement with a previous report[20] most misclassifications occurred in the borderline group. One in three borderline masses was misclassified as benign by pattern recognition and the scoring system, and one in five by the LR model.

A previous study on the use of pattern recognition for discrimination between benign and malignant adnexal masses showed that non-expert ultrasound operators reached a sensitivity and specificity with regard to malignancy of 86% and 80%, respectively[5]. In the current study, where the scans were carried out and interpreted by very experienced examiners, pattern recognition had a sensitivity of 90% and a specificity of 93% in the prospective validation test. The sensitivity and specificity of our subgroup scoring system (sensitivity 88%, specificity 90% in the prospective validation set) compare favorably with those of pattern recognition. Pattern recognition in the hands of experienced examiners has been shown to be the best diagnostic method currently available for discrimination between benign and malignant masses[17,21–23]. In the prospective validation set, the subgroup scoring system resulted in a substantial improvement in specificity at the cost of a small reduction in sensitivity when compared with our overall LR model.

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

Table 4 Sensitivity and specificity with regard to malignancy of the overall logistic regression (LR) model, subjective evaluation of the ultrasound image (pattern recognition) by experienced ultrasound examiners and the subgroup scoring system, results being stratified by subgroups of tumor

	Development set			*Test set*			*Prospective validation set*		
Subgroup	*Overall LR model*	*Pattern recognition*	*Subgroup scoring system*	*Overall LR model*	*Pattern recognition*	*Subgroup scoring system*	*Overall LR model*	*Pattern recognition*	*Subgroup scoring system*
Unilocular									
Sensitivity	0 (0/1)	0 (0/1)	0 (0/1)	0 (0/1)	0 (0/1)	0 (0/1)	0 (0/2)	0 (0/2)	0 (0/2)
Specificity	98 (222/227)	99.6 (226/227)	100 (227/227)	99 (83/84)	100 (84/84)	100 (84/84)	94 (134/142)	100 (142/142)	100 (142/142)
Correct	97 (222/228)	99 (226/228)	99.6 (227/228)	98 (83/85)	99 (84/85)	99 (84/85)	93 (134/144)	99 (142/144)	99 (142/144)
LR+	0	0	—	0	—	—	0	—	—
LR−	1.02	1.00	1	1.01	1	1	1.06	1	1
Multilocular									
Sensitivity	40 (6/15)	40 (6/15)	53 (8/15)	40 (2/5)	60 (3/5)	40 (2/5)	33 (1/3)	67 (2/3)	67 (2/3)
Specificity	93 (119/128)	98 (126/128)	95 (122/128)	88 (46/52)	98 (51/52)	90 (47/52)	87 (68/78)	95 (74/78)	92 (72/78)
Correct	87 (125/143)	92 (132/143)	91 (130/143)	84 (48/57)	95 (54/57)	86 (49/57)	85 (69/81)	94 (76/81)	91 (74/81)
LR+	5.69	25.60	11.38	3.47	31.20	4.16	2.60	13	8.67
LR−	0.65	0.61	0.49	0.68	0.41	0.66	0.76	0.35	0.35
Solid component, no papillation									
Sensitivity	98 (84/86)	95 (82/86)	88 (76/86)	100 (37/37)	89 (33/37)	92 (34/37)	97 (67/69)	87 (60/69)	84 (58/69)
Specificity	54 (52/97)	95 (92/97)	81 (79/97)	54 (30/56)	93 (52/56)	88 (49/56)	46 (34/74)	88 (65/74)	85 (63/74)
Correct	74 (136/183)	95 (174/183)	85 (155/183)	72 (67/93)	91 (85/93)	89 (83/93)	71 (101/143)	87 (125/143)	85 (121/143)
LR+	2.11	18.50	4.76	2.15	12.49	7.35	1.80	7.15	5.65
LR−	0.04	0.05	0.14	0	0.12	0.09	0.06	0.15	0.19
Papillation									
Sensitivity	98 (87/89)	91 (81/89)	93 (83/89)	97 (31/32)	91 (29/32)	88 (28/32)	99 (68/69)	97 (67/69)	96 (66/69)
Specificity	36 (40/111)	86 (96/111)	75 (83/111)	44 (20/45)	78 (35/45)	69 (31/45)	49 (34/70)	81 (57/70)	74 (52/70)
Correct	64 (127/200)	89 (177/200)	83 (166/200)	66 (51/77)	83 (64/77)	77 (59/77)	73 (102/139)	89 (124/139)	85 (118/139)
LR+	1.53	6.73	3.70	1.74	4.08	2.81	1.92	5.23	3.72
LR−	0.06	0.10	0.09	0.07	0.12	0.18	0.03	0.04	0.06
Total									
Sensitivity	93 (177/191)	88 (169/191)	87 (167/191)	93 (70/75)	87 (65/75)	85 (64/75)	95 (136/143)	90 (129/143)	88 (126/143)
Specificity	77 (433/563)	96 (540/563)	91 (511/563)	76 (179/237)	94 (222/237)	89 (211/237)	74 (270/364)	93 (338/364)	90 (329/364)
Correct	81 (610/754)	94 (709/754)	90 (678/754)	80 (249/312)	92 (287/312)	88 (275/312)	80 (406/507)	92 (467/507)	90 (455/507)
LR+	4.01	21.66	9.47	3.81	13.69	7.78	3.68	12.63	9.16
LR−	0.10	0.12	0.14	0.09	0.14	0.16	0.07	0.11	0.13

Sensitivity, Specificity and Correct are presented as % (*n*).

Table 5 The ability of the overall logistic regression (LR) model, pattern recognition (subjective evaluation of the ultrasound image) and the subgroup scoring system to correctly classify a mass as benign or malignant depending on its specific histology ($n = 1573$ patients)

Parameter	*Correctly classified (n (%))*		
	Overall LR model	*Pattern recognition*	*Subgroup scoring system*
Benign pathology			
Endometrioma ($N = 313$)	282 (90)	308 (98)	300 (96)
Dermoid/teratoma ($N = 176$)	153 (87)	175 (99)	168 (95)
Simple cyst ($N = 189$)*	159 (84)	187 (99)	180 (95)
Hydrosalpinx ($N = 51$)†	35 (69)	47 (92)	43 (84)
Peritoneal pseudocyst ($N = 10$)	8 (80)	10 (100)	10 (100)
Abscess ($N = 18$)	4 (22)	15 (83)	14 (78)
Fibroma ($N = 66$)‡	15 (23)	57 (86)	41 (62)
Serous cystadenoma ($N = 184$)	122 (66)	164 (89)	158 (86)
Mucinous cystadenoma ($N = 132$)	95 (72)	121 (92)	119 (90)
Rare benign ($N = 25$)§	9 (36)	16 (64)	18 (72)
Malignant pathology			
Primary invasive ($N = 272$)	263 (97)	254 (93)	248 (91)
Stage I ($N = 66$)	61 (92)	58 (88)	56 (85)
Stage II ($N = 17$)	17 (100)	17 (100)	14 (82)
Stage III ($N = 132$)	130 (98)	126 (95)	127 (96)
Stage IV ($N = 28$)	28 (100)	28 (100)	27 (96)
Rare malignant ($N = 29$)¶	27 (93)	25 (86)	24 (83)
Borderline ($N = 75$)	59 (79)	50 (67)	52 (69)
Stage I ($N = 65$)	49 (75)	40 (62)	43 (66)
Stage II ($N = 5$)	5 (100)	5 (100)	5 (100)
Stage III ($N = 5$)	5 (100)	5 (100)	4 (80)
Metastatic ($N = 62$)	61 (98)	59 (95)	57 (92)

*Including parasalpingeal cyst, inclusion cyst, functional cyst and normal ovary. †Including salpingitis. ‡Including one case of ovarian leiomyoma and twelve cases of uterine leiomyoma or lipoleiomyoma, the latter having been misdiagnosed as an ovarian mass by the ultrasound examiner. §Including struma ovarii, Brenner tumor, Sertoli cell tumor, stromal cell tumor, Schwannoma and lymphangioma. ¶Including granulosa cell tumor, Leydig cell tumor, dysgerminoma, gynandroblastoma, leiomyosarcoma, immature teratoma, malignant mixed Müllerian tumor, small cell cancer, Brenner cancer, carcinosarcoma, choriocarcinoma and yolk sac tumor. *N*, total number with the condition; *n*, number in the subgroup.

Table 6 Prevalence of benign and malignant tumors for possible values of the scoring system, results being stratified by subgroups of tumors

Parameter	n	*Benign (n (%))*	*Malignant (n (%))*			
			Total	*Borderline*	*Primary invasive*	*Metastatic*
Unilocular tumors	457	453 (99)	4 (< 1)	3 (< 1)	1 (< 1)	—
Scoring system for multilocular tumors						
Total score						
0	98	98 (100)	—	—	—	—
1	104	98 (94)	6 (6)	1 (1)	5 (5)	—
2	50	45 (90)	5 (10)	4 (8)	1 (2)	—
3	23	16 (70)	7 (30)	5 (22)	2 (9)	—
4	3	1 (33)	2 (67)	—	2 (67)	—
5	3	—	3 (100)	1 (33)	1 (33)	1 (33)
Scoring system for tumors with a solid component but no papillation						
Total score						
≤ 1	155	143 (92)	12 (8)	4 (3)	5 (3)	3 (2)
2	23	19 (83)	4 (17)	2 (9)	1 (4)	1 (4)
3	37	29 (78)	8 (22)	1 (3)	6 (16)	1 (3)
4	17	10 (59)	7 (41)	1 (6)	5 (29)	1 (6)
5	30	11 (37)	19 (63)	2 (7)	10 (33)	7 (23)
≥ 6	157	15 (10)	142 (90)	4 (3)	104 (66)	34 (22)
Scoring system for tumors with papillation						
Total score						
≤ −1	107	104 (97)	3 (3)	3 (3)	—	—
0	41	36 (88)	5 (12)	4 (10)	1 (2)	—
1	31	26 (84)	5 (16)	1 (3)	4 (13)	—
2	34	22 (65)	12 (35)	6 (18)	5 (15)	1 (3)
3	24	14 (58)	10 (42)	3 (13)	6 (25)	1 (4)
≥ 4	179	24 (13)	155 (87)	30 (17)	113 (63)	12 (7)

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

In addition to having better diagnostic performance than the overall LR model, our subgroup scoring system is more user-friendly (Figure 3). Fewer variables are needed to classify a tumor as benign or malignant, e.g. for multilocular masses only four characteristics are needed vs. 12 for the overall LR model. Moreover, only a piece of paper is needed to note the scores while a computer is needed to calculate the risk of malignancy using the LR model. Both the overall LR model and our subgroup scoring system can be used to estimate the risk of malignancy on a sliding scale. The overall LR model gives the risk on a scale from 0 to 100%, while a risk estimate based on the subgroup scoring system is rougher and can be summarized as 'the higher the score the higher the risk' (Table 6).

Unfortunately, it is not possible to recommend clinical management on the basis of an estimated risk of malignancy, because appropriate management of patients is determined by many factors, e.g. symptoms, operative risk, anxiety level and personal preferences of patients, but an estimated risk of malignancy based on the overall LR model or the subgroup scoring system can be of help to the clinician when deciding on which treatment to recommend to the patient. However, before the subgroup scoring system is introduced into clinical practice it needs to be tested prospectively outside the IOTA centers, where the scoring system was created, and also by non-expert ultrasound examiners. This is clearly very important, not least because the subgroup scoring system – as well as the overall LR model and pattern recognition – uses ultrasound variables that are based on subjective evaluation, e.g. evaluation of regularity of cyst walls and estimation of the color content of the tumor scan (color score). Even deciding on whether or not a tumor contains a solid component is based on subjective evaluation. This element of subjectivity might have a negative effect on the reliability of the subgroup scoring system (as well as on that of the other ultrasound methods). To the best of our knowledge, intraobserver repeatability and interobserver agreement with regard to 'irregularity in a tumor' or the presence or absence of solid components in a tumor have not been determined, but subjective estimation of the color content of a tumor scan on a visual analog scale has been found to be reproducible enough for clinical use[24].

The robustness of our subgroup scoring system may be limited by the development sets of the different subgroup scores being rather small compared to the development set of the overall LR model (143 to 228 cases vs. 754). Nevertheless, the variables in the subgroup scoring system were selected in order to fit with the level of clinical expertise, e.g. ascites is an important variable in each subgroup. Some fine-tuning of the subgroup scoring system might be needed. This will be addressed in phases 2 and 3 of the IOTA study.

In conclusion, our subgroup scoring system is a relatively simple method that should be tested in less experienced hands to see if its use will improve the characterization of ovarian tumors by medical professionals that are not experts in gynecological ultrasonography.

ACKNOWLEDGMENTS

This research was supported by GOA-AMBioRICS, CoE EF/05/006, Belgian network DYSCO, BIOPATTERN (FP6-2002-IST 508803), ETUMOUR (FP6-2002-LIFESCIHEALTH 503094), IWT-TBM 070706 (IOTA), Swedish Medical Research Council (grants nos. K2001-72X-11605-06A, K2002-72X-11605-07B, K2004-73X-11605-09A and 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); ALF- medel (a Swedish governmental grant from the region of Scania).

REFERENCES

1. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ. Cancer statistics, 2007. *CA Cancer J Clin* 2007; 57: 43–66.
2. Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P, Gore ME, Kaern J, Verrelst H, Sjovall K, Timmerman D, Vandewalle J, Van Gramberen M, Tropé CG. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. *Lancet* 2001; 357: 176–182.
3. 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.
4. Granberg S, Wikland M, Jansson I. Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis. Criteria to be used for ultrasound evaluation. *Gynecol Oncol* 1989; 35: 139–144.
5. Valentin L. Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. *Ultrasound Obstet Gynecol* 1999; 15: 338–347.
6. Timmerman D, Schwärzler 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. 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.
8. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol* 1990; 97: 922–929.
9. 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.
10. Tailor A, Jurkovic D, Bourne TH, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. *Ultrasound Obstet Gynecol* 1997; 10: 41–47.
11. Timmerman D, Bourne T, Tailor A, Collins WP, Verrelst H, Vandenberghe K, Vergote I. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model. *Am J Obstet Gynecol* 1999; 181: 57–65.
12. Timmerman D, Verrelst H, Bourne TH, De Moor B, Collins WP, Vergote I, Vandewalle J. Artificial neural network models for the preoperative discrimination between malignant and

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.

benign adnexal masses. *Ultrasound Obstet Gynecol* 1999; 13: 17–25.
13. Tailor A, Jurkovic D, Bourne TH, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using an artificial neural network. *Br J Obstet Gynaecol* 1999; 106: 21–30.
14. Timmerman D. Lack of standardization in gynecological ultrasonography (editorial). *Ultrasound Obstet Gynecol* 2000; 16: 395–398.
15. Aslam N, Banerjee S, Carr JV, Savvas M, Hooper R, Jurkovic D. Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer. *Obstet Gynecol* 2000; 96: 75–80.
16. Mol BWJ, Boll D, De Kanter M, Heintz APM, Sijmons EA, Oei SG, Bal H, Brölmann HAM. Distinguishing the benign and malignant adnexal mass: An external validation of prognostic models. *Gynecol Oncol* 2001; 80: 162–167.
17. 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.
18. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I; International Ovarian Tumor Analysis (IOTA) Group. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: A consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound Obstet Gynecol* 2000; 16: 500–505.
19. 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.
20. 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.
21. 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.
22. 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* (in press).
23. 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.
24. Sladkevicius P, Valentin L. Interobserver agreement in the results of Doppler examinations of extrauterine pelvic tumors. *Ultrasound Obstet Gynecol* 1995; 6: 91–96.

Copyright © 2008 ISUOG. Published by John Wiley & Sons, Ltd.