Ultrasound Obstet Gynecol 2011; **38**: 456–465
Published online 13 September 2011 in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.9030

Adnexal masses difficult to classify as benign or malignant using subjective assessment of gray-scale and Doppler ultrasound findings: logistic regression models do not help

L. VALENTIN*, L. AMEYE†‡, L. SAVELLI§, R. FRUSCIO¶, F. P. G. LEONE**, A. CZEKIERDOWSKI††, A. A. LISSONI¶, D. FISCHEROVA‡‡, S. GUERRIERO§§, C. VAN HOLSBEKE¶¶ #, S. VAN HUFFEL†‡ and D. TIMMERMAN#

**Department of Obstetrics and Gynecology, Skåne University Hospital Malmö, Lund University, Malmö, Sweden; †Department of Electrical Engineering, ESAT-SCD(SISTA), Katholieke Universiteit Leuven, Leuven, Belgium; ‡IBBT-K.U.Leuven Future Health Department, Leuven, Belgium; §Gynecology and Reproductive Medicine Unit, Department of Obstetrics and Gynecology, University of Bologna, Bologna, Italy; ¶Clinica Ostetrica e Ginecologica, Ospedale S. Gerardo, Università di Milano Bicocca, Monza, Italy; **Department of Obstetrics and Gynecology, Clinical Sciences Institute L. Sacco, University of Milan, Milan, Italy; ††1st Department of Gynecologic Oncology and Gynecology, Medical University in Lublin, Lublin, Poland; ‡‡Gynecological Oncology Centre, Department of Obstetrics and Gynecology, First Faculty of Medicine and General University Hospital, Charles University, Prague, Czech Republic; §§Department of Obstetrics and Gynecology of the University of Cagliari, Ospedale San Giovanni di Dio, Cagliari, Italy; ¶¶Department of Obstetrics and Gynecology, Ziekenhuis Oost-Limburg, Genk, Belgium; #Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Leuven, Belgium*

KEYWORDS: CA 125 antigen; logistic models; ovarian neoplasms; sensitivity; specificity; ultrasonography

ABSTRACT

Objective *To develop a logistic regression model that can discriminate between benign and malignant adnexal masses perceived to be difficult to classify by subjective evaluation of gray-scale and Doppler ultrasound findings (subjective assessment) and to compare its diagnostic performance with that of subjective assessment, serum CA 125 and the risk of malignancy index (RMI).*

Methods *We used data from the 3511 patients with an adnexal mass included in the International Ovarian Tumor Analysis (IOTA) studies. All patients had been examined using transvaginal gray-scale and Doppler ultrasound following a standardized research protocol carried out by an experienced ultrasound examiner using a high-end ultrasound system. In addition to prospectively collecting information on > 40 clinical and ultrasound variables, the ultrasound examiner classified each mass as certainly or probably benign, unclassifiable, or certainly or probably malignant. A logistic regression model to discriminate between benignity and malignancy was developed for the unclassifiable masses (*n = 244*, i.e. 7% of all tumors) using a training set (160 tumors, 45 malignancies) and then tested on a test set (84 tumors, 28 malignancies). The gold standard was the histological diagnosis of the surgically removed adnexal mass. The area under the receiver–operating characteristics curve (AUC), sensitivity, specificity, positive likelihood ratio (LR+) and negative likelihood ratio (LR−) were used to describe diagnostic performance and were compared between subjective assessment, CA 125, the RMI and the logistic regression model created.*

Results *One variable was retained in the logistic regression model: the largest diameter (in mm) of the largest solid component of the tumor (odds ratio (OR) = 1.04; 95% CI, 1.02–1.06). The model had an AUC of 0.68 (95% CI, 0.59–0.78) on the training set and an AUC of 0.65 (95% CI, 0.53–0.78) on the test set. On the test set, a cut-off of 25% probability of malignancy (corresponding to the largest diameter of the largest solid component of 23 mm) resulted in a sensitivity of 64% (18/28), a specificity of 55% (31/56), an LR+ of 1.44 and an LR− of 0.65. The corresponding values for subjective assessment were 68% (19/28), 59% (33/56), 1.65 and 0.55. On the test set of patients with available CA 125 results, the LR+ and LR− of the logistic regression model (cut-off = 25% probability of malignancy) were 1.29 and 0.73, of subjective assessment were 1.45 and 0.63, of CA 125 (cut-off = 35 U/mL) were 1.24 and 0.84 and of RMI (cut-off = 200) were 1.21 and 0.92.*

Conclusions *About 7% of adnexal masses that are considered appropriate for surgical removal cannot be classified as benign or malignant by experienced ultrasound*

Correspondence to: Prof. L. Valentin, Department of Obstetrics and Gynecology, Skåne Hospital Malmö, SE 205 02 Malmö, Sweden (e-mail: lil.valentin@med.lu.se)

Accepted: 8 April 2011

examiners using subjective assessment. Logistic regression models to estimate the risk of malignancy, CA 125 measurements and the RMI are not helpful in these masses.

INTRODUCTION

To be able to offer women with an adnexal mass optimal treatment, one needs to know whether the mass is likely to be benign or malignant. If surgery is required, the method of surgery will depend on the nature of the mass. Most benign cysts can be treated with minimally invasive surgery, which is associated with a shorter duration of hospital stay and rehabilitation than is laparotomy[1,2]. Patients with a malignant tumor need to undergo extensive staging procedures[3]. Performing laparoscopic surgery, even in early stage ovarian cancer, should be avoided because rupture during surgery on a Stage I ovarian cancer may worsen the prognosis[4].

One of the best methods for discriminating between benign and malignant adnexal masses is subjective assessment (i.e. subjective evaluation of gray-scale and Doppler ultrasound findings by an experienced ultrasound examiner[5,6]; also called pattern recognition). However, using subjective assessment, a small proportion of masses cannot be confidently classified as benign or malignant ('unclassifiable masses')[7]. For such masses, methods other than subjective assessment are needed. Other methods to classify adnexal masses as benign or malignant are measurement of the serum CA 125 level, calculation of the risk of malignancy index (RMI)[8] and the use of various mathematical models to calculate the risk of malignancy[9]. In a previous study, we were unable to construct a logistic regression model to estimate the risk of malignancy in unclassifiable masses[7]. That study included 90 unclassifiable masses, and no clinical or ultrasound variable was retained in a logistic regression model to predict malignancy in these masses.

The aim of this study was to develop a logistic regression model that can discriminate between benign and malignant unclassifiable adnexal masses in a larger study population than in our previous study. We also wanted to compare the diagnostic performance of the model developed with that of subjective assessment, CA 125 and the RMI.

METHODS

We used data from the 3511 patients with an adnexal mass in the International Ovarian Tumor Analysis (IOTA) studies Phase 1[10], Phase 1b[11] and Phase 2[12]. The IOTA study is a prospective international multicenter study of patients with adnexal masses. Patients were recruited in 21 ultrasound centers in nine countries between 1999 and 2007. The centers and the principal investigators at each center are listed in the Appendix. The research protocol was ratified by the local Ethics Committee at each center. Detailed information on the IOTA studies and the IOTA ultrasound protocol can be found in the literature[10–13]. A short description is given below.

The patients included were scanned by a gynecologist or radiologist with expertise in gynecological ultrasound and a special interest in adnexal pathology (i.e. a principal investigator at each participating center). Because the principal investigator was not always available to perform the scan, our study population is not a strictly consecutive series of patients. Patients who were pregnant or refused transvaginal ultrasound were not eligible for inclusion, and patients who did not undergo surgical removal of the mass within 120 days after the ultrasound examination were excluded. The decision to operate or not was made by local clinicians on the basis of the results of the ultrasound examination (subjective assessment), the clinical picture and the local management protocols.

The patients were scanned transvaginally using high-end ultrasound systems equipped with high-frequency transvaginal probes. A dedicated research protocol was followed, and the IOTA terms and definitions were used to describe the ultrasound findings[13]. Transabdominal ultrasound was added if the mass was so large that it could not be seen in its entirety using a transvaginal probe. In addition to prospectively collecting information on more than 40 ultrasound variables and a few clinical variables, at the end of the ultrasound examination the ultrasound examiner classified each mass as benign or malignant using subjective assessment (also called pattern recognition). Moreover, he/she reported the level of diagnostic confidence with which the prediction of benignity/malignancy was made: certainly benign, probably benign, unclassifiable, probably malignant or certainly malignant. Even when the ultrasound examiner found the mass impossible to classify (unclassifiable mass), he/she was obliged to classify it as most likely benign or most likely malignant. In the case of bilateral adnexal masses, the mass with the most complex ultrasound morphology was included in our statistical analysis. If both masses had similar ultrasound morphology, the largest one or the one most easily accessible by transvaginal ultrasound was included. CA 125 results were not available to the ultrasound examiner at the time of the ultrasound examination.

The prospectively collected clinical and ultrasound information was recorded in an electronic data-collection system. The information was locked at the time of the examination and could not be changed thereafter. Logistic regression models to calculate the risk of malignancy were created after completion of patient recruitment and played no role in the management of the patients.

The participating centers were encouraged to measure the level of serum CA 125 in peripheral blood from all patients, but the availability of CA 125 results was not a requirement for inclusion in the IOTA studies. Second-generation immunoradiometric assay kits for CA 125 (i.e. CA 125II)[14] from six companies were used (Roche Diagnostics, Basel, Switzerland; Centocor, Malvern, PA, USA; Cis-Bio, Gif-sur-Yvette, France; Abbott Laboratories Diagnostic Division, Abbott Park, IL, USA; Bayer Diagnostics, Tarrytown, NY, USA; and bioMérieux, Marcy

l'Etoile, France). All kits used the OC 125 antibody. CA 125 results are expressed in units per milliliter (U/mL).

The gold standard was the histological diagnosis of the surgically removed adnexal mass. The excised tissues underwent histological examination at the local center. Tumors were classified and staged (if malignant) according to the criteria recommended by the International Federation of Gynecology and Obstetrics[15,16].

Statistical analysis

Statistical analyses were carried out using Statistical Analysis Software (SAS) version 9.2 (SAS Institute Inc., Cary, NC, USA).

The Student's *t*-test and the Mann–Whitney *U*-test were used to determine the statistical significance of differences in continuous data, and the chi-square test and Fisher's exact test were used to determine the statistical significance of differences in unpaired discrete data. The permutation method[17] was used to correct for multiple testing. To test the statistical significance of differences in paired discrete data (sensitivity and specificity) we used the McNemar test. Exact 95% CI values for sensitivity and specificity were calculated using the Clopper–Pearson method, and 95% CI values for the positive likelihood ratio (LR+) and the negative likelihood ratio (LR−) were calculated using the method described by Simel *et al.*[18].

Logistic regression with stepwise selection of variables was used (i) to determine which ultrasound variables were independently associated with an unclassifiable mass and (ii) for building a model to predict malignancy in unclassifiable masses. In the model-building process, structural missing values for information about papillary projections and solid components were substituted with zero (e.g. if there was no papillation, the height of the largest papillary structure was imputed with a zero). When intratumoral arterial blood flow velocity waveforms were not detected, the peak systolic velocity, time-averaged maximum velocity, pulsatility index (PI) and resistance index (RI) were coded as 2.0 cm/s, 1 cm/s, 3.0 and 1.0, respectively, for use in mathematical modeling[10].

For building a logistic regression model to calculate the risk of malignancy in unclassifiable adnexal masses, the whole IOTA database was divided into a training set comprising 70% of the tumors ($n = 2469$) and a test set comprising 30% of the tumors ($n = 1042$). When

Table 1 Histological diagnosis for classifiable and unclassifiable (difficult) masses

Histological diagnosis	*Classifiable masses* (n = *3267*)	*Unclassifiable masses* (n = *244*)	P*	*All unclassifiable masses* (n = *244*)		*Unclassifiable masses with available CA 125 values* (n = *191*)	
				Training set (n = *160*)	*Test set* (n = *84*)	*Training set* (n = *123*)	*Test set* (n = *68*)
Benign	2389 (73.1)	171 (70.1)		115 (71.9)	56 (66.7)	85 (69.1)	42 (61.8)
Endometrioma	691 (21.2)	22 (9.0)	< 0.001	14 (8.8)	8 (9.5)	11 (8.9)	5 (7.4)
Teratoma	391 (12.0)	11 (4.5)	0.001	8 (5.0)	3 (3.6)	7 (5.7)	2 (2.9)
Simple cyst + parasalpingeal cyst	269 (8.2)	12 (4.9)	0.37	8 (5.0)	4 (4.7)	6 (4.9)	4 (5.9)
Functional cyst	110 (3.4)	6 (2.5)		4 (2.5)	2 (2.4)	1 (0.8)	2 (2.9)
Hydrosalpinx + salpingitis	96 (2.9)	4 (1.6)		2 (1.3)	2 (2.4)	2 (1.6)	2 (2.9)
Peritoneal pseudocyst	21 (0.6)	0 (0)		0 (0)	0 (0)	0 (0)	0 (0)
Abscess	37 (1.1)	5 (2.1)	0.83	4 (2.5)	1 (1.2)	3 (2.4)	1 (1.5)
Fibroma	130 (4.0)	22 (9.0)	0.006	17 (10.6)	5 (6.0)	14 (11.4)	3 (4.4)
Serous cystadenoma	370 (11.3)	50 (20.5)	< 0.001	33 (19.4)	17 (20.2)	24 (19.5)	11 (16.2)
Mucinous cystadenoma	239 (7.3)	31 (12.7)	< 0.001	21 (13.1)	10 (11.9)	14 (11.4)	9 (13.2)
Rare benign†	35 (1.1)	8 (3.3)	0.06	4 (2.5)	4 (4.7)	3 (2.4)	3 (4.4)
Borderline	153 (4.7)	33 (13.5)	< 0.001	21 (13.1)	12 (14.3)	19 (15.4)	12 (17.6)
Borderline Stage I	132 (4.0)	32 (13.1)		20 (12.5)	12 (14.3)	18 (14.6)	12 (17.6)
Borderline Stage II	7 (0.2)	1 (0.4)		1 (0.6)	0 (0)	1 (0.8)	0 (0)
Borderline Stage III	13 (0.4)	0 (0)		0 (0)	0 (0)	0 (0)	0 (0)
Borderline Stage IV	1 (0.03)	0 (0)		0 (0)	0 (0)	0 (0)	0 (0)
Primary invasive	615 (18.8)	30 (12.3)		19 (11.9)	11 (13.1)	14 (11.4)	11 (16.2)
Primary invasive Stage I	128 (3.9)	8 (3.3)		5 (3.1)	3 (3.6)	3 (2.4)	3 (4.4)
Primary invasive Stage II	46 (1.4)	1 (0.4)		1 (0.6)	0 (0)	0 (0)	0 (0)
Primary invasive Stage III	320 (9.8)	14 (5.7)		8 (5.0)	6 (7.1)	7 (5.7)	6 (8.8)
Primary invasive Stage IV	57 (1.7)	1 (0.4)		1 (0.6)	0 (0)	1 (0.8)	0 (0)
Rare primary invasive‡	64 (2.0)	6 (2.5)		4 (2.5)	2 (2.4)	3 (2.4)	2 (2.9)
Metastatic	110 (3.4)	10 (4.1)		5 (3.1)	5 (6.0)	5 (4.0)	3 (4.4)
All invasive	725 (22.2)	40 (16.4)	0.07	24 (15.0)	16 (19.0)	19 (15.4)	14 (20.6)

Data are given as *n* (%). **P*-values (Fisher's exact test) were corrected for multiple testing using the permutation method[17]. †Struma ovarii, Brenner tumor, Leydig cell tumor, stromal cell tumor, steroid cell tumor of stromal luteoma variant, Schwannoma, lymphangioma, cystic mesothelioma, leiomyoma of bowel, gynandroblastoma. ‡Sertoli cell tumor, Sertoli-Leydig cell tumor, granulosa cell tumor, carcinosarcoma, dysgerminoma, immature teratoma and others.

creating the training and test sets, the data were stratified for outcome (benign, borderline, primary invasive or metastatic tumor) and center to ensure that the proportion of malignant and benign masses and the proportion of masses derived from each contributing center were the same for the development set and the test set. The training set for unclassifiable masses comprised the 160 unclassifiable masses in the whole training set. The test set for the unclassifiable masses comprised the 84 unclassifiable masses in the whole test set. Of the 160 unclassifiable masses in the training set, 123 had information on CA 125. Of the 84 unclassifiable masses in the test set, 68 had information on CA 125. The 68 masses in the test set with available CA 125 results were used to compare the diagnostic performance of subjective assessment, the logistic regression model created, RMI and CA 125. A receiver–operating characteristics (ROC) curve was drawn to evaluate the diagnostic ability of the model and to determine the mathematically best cut-off value to predict malignancy, the mathematically best cut-off value being defined as that corresponding to the point on the ROC curve situated farthest from the reference line.

The diagnostic performance of the logistic regression model, CA 125 and the RMI was expressed as the area under the ROC curve (AUC). If the lower limit of the 95% CI for the AUC was > 0.5, the test was considered to have discriminatory potential. In addition, the sensitivity, specificity, LR+ and LR− were calculated for subjective assessment (using the examiner's dichotomous classification of the mass as benign or malignant), the logistic regression model (using the mathematically best cut-off of the model, or a cut-off that yielded the same sensitivity as subjective assessment in the training set), CA 125 (using a cut-off of 35 U/mL[19]) and RMI (using a cut-off of 200[8]). The statistical significance of a difference in AUC was determined using the technique described by de Long *et al.*[20].

RESULTS

The ultrasound examiner was uncertain whether the adnexal mass was benign or malignant in 244 (7.0%; 99% CI, 5.8–8.1) of the 3511 masses. Forty (16%) of the 244 unclassifiable masses were invasive malignancies and 33 (13.5%) were borderline malignancies. The histology of the masses in the training sets and test sets is shown in Table 1. Borderline tumors, fibromas, and serous and mucinous cystadenomas/cystadenofibromas were two to three times more common among the unclassifiable masses than among the classifiable masses. Of the 1028 borderline tumors, fibromas, cystadenomas and cystadenofibromas, 13% (136/1028) were unclassifiable, 9% (93/1028) were incorrectly classified as benign or malignant and 78% (799/1028) were correctly classified by the ultrasound examiner using subjective assessment. Of the remaining 2483 tumors, 3% (80/2483) were unclassifiable, 4% (108/2483) were incorrectly classified as benign or malignant and 92% (2295/2483) were correctly classified by subjective assessment. The difference was statistically significant ($P < 0.001$). Ultrasound images of unclassifiable tumors are shown in Figures 1–3.

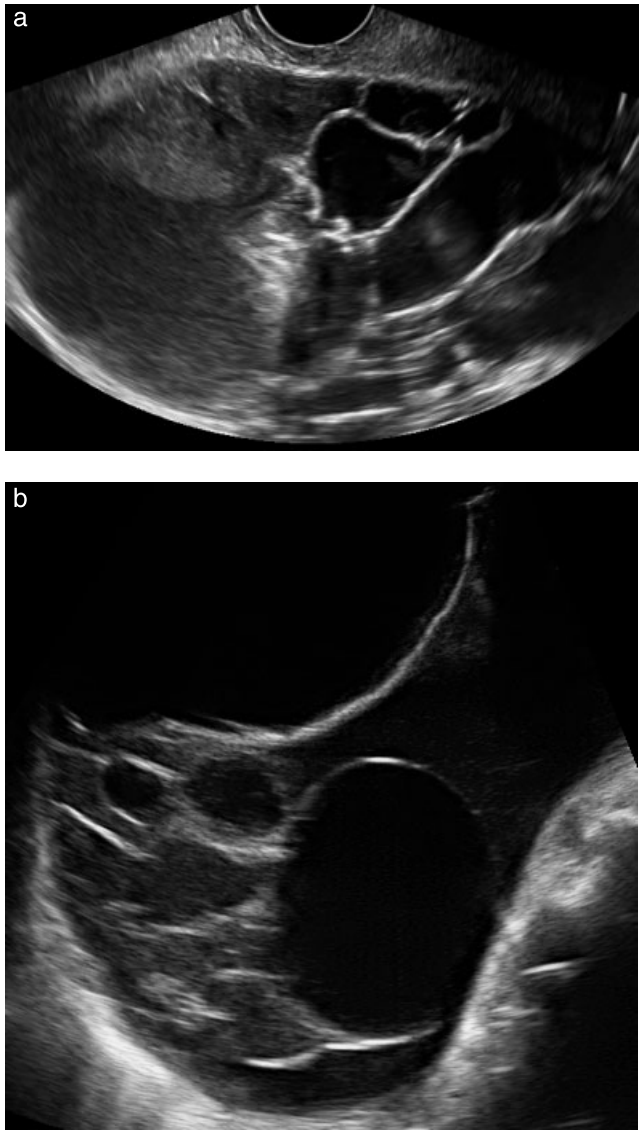

Figure 1 Two unclassifiable multilocular masses with more than 10 cyst locules: a 13-cm benign mucinous cystadenoma in an 81-year-old woman (a) and a 17-cm mucinous borderline tumor in a 63-year-old woman (b).

Clinical and ultrasound information for the classifiable and unclassifiable masses is shown in Table S1. The unclassifiable masses were larger than the classifiable masses, they were more often unilocular solid or multilocular solid masses (64% vs. 31%, $P < 0.001$) and they more often had irregular walls (65% vs. 39%, $P < 0.001$) and papillary projections (48% vs. 20%, $P < 0.001$) than classifiable masses, but had fewer papillary projections and smaller solid components. Multilocular cysts with more than 10 cyst locules were also more common among the unclassifiable masses (5% vs. 2%, $P = 0.005$). Absence of color Doppler signals was less common (16% vs. 30%) in the unclassifiable masses, while a moderate amount of color Doppler signals (color score 3) was more common (37% vs. 27%). Women with an unclassifiable mass were older and of higher parity

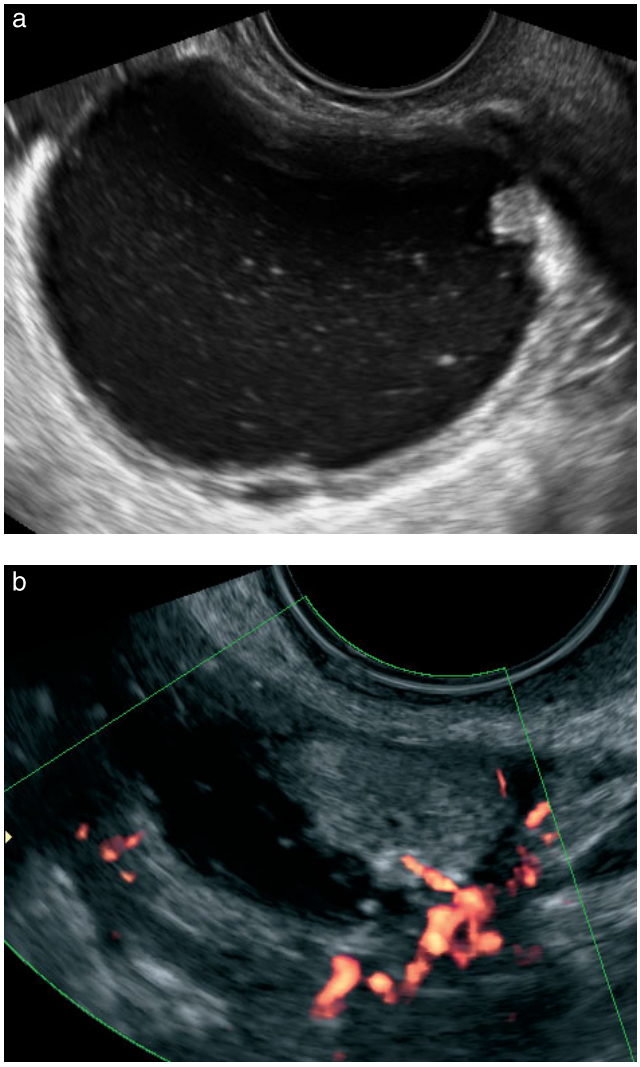

Figure 2 An unclassifiable 6-cm mass with papillary projections in a 34-year-old woman. The histopathological diagnosis was benign mucinous cyst. At gray-scale ultrasound examination a few papillary projections were seen, one of which is shown (a), and at color Doppler examination a vessel was seen to penetrate into the papillation from the cyst wall (b).

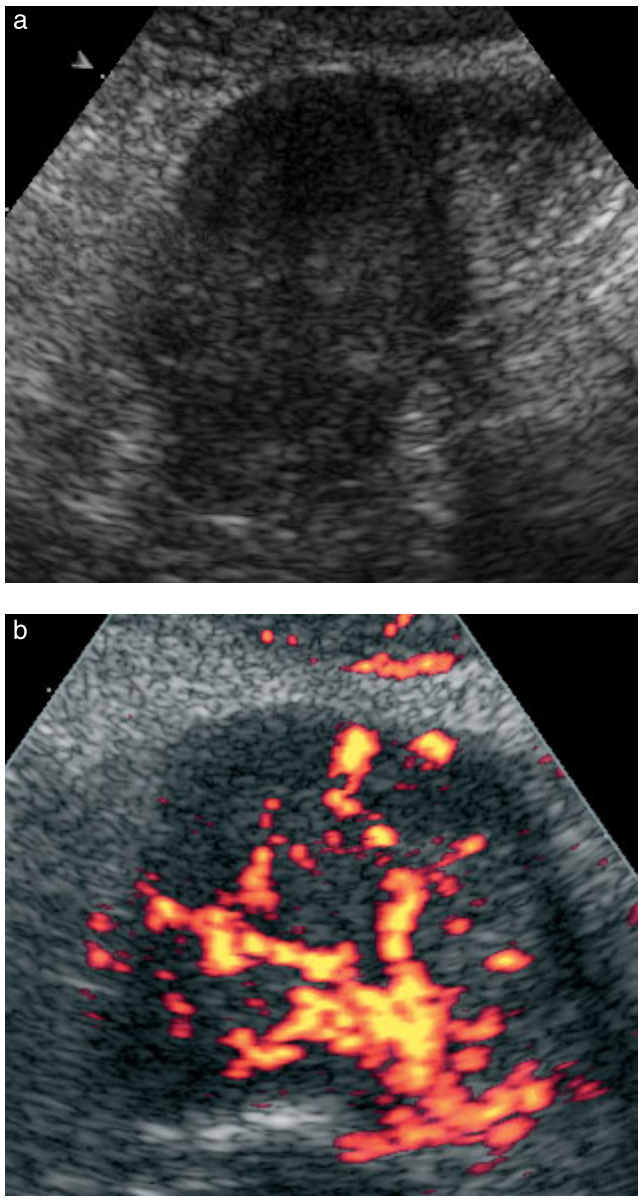

Figure 3 An unclassifiable 2.5-cm solid tumor in a 61-year-old woman with postmenopausal bleeding. The histopathological diagnosis was ovary with stromal hyperplasia. The gray-scale image revealed slightly irregular internal echogenicity (a) and the mass was highly vascularized at color Doppler examination (b).

than those with a classifiable mass, more had undergone hysterectomy and more had a personal history of ovarian cancer. Multiple logistic regression analysis showed the following variables to be independently associated with an unclassifiable mass: menopausal status (odds ratio (OR) = 1.37; 95% CI, 1.04–1.80), unilocular cyst (OR = 0.26; 95% CI, 0.12–0.55), multilocular cyst with more than 10 cyst locules (OR = 3.84; 95% CI, 1.78–8.30), low-level echogenicity of cyst fluid (OR = 2.02; 95% CI, 1.50–2.72), presence of solid components (OR = 5.78; 95% CI, 3.38–9.87) and the largest diameter of the largest solid component (OR = 0.98 per 1-mm increase; 95% CI, 0.97–0.99).

Ultrasound findings and clinical background data in benign and malignant unclassifiable masses are shown in Table 2. The malignant unclassifiable masses were larger than the benign masses, and solid components were more common in the malignant masses (89% vs. 80%), but these differences did not reach statistical significance ($P = 0.06$ for both comparisons). In masses with solid components, the solid components were larger in the malignant masses than in the benign masses. Papillary projections were equally common in benign and malignant unclassifiable masses, but if papillary projections were present, there were more papillary projections in the malignant masses than in the benign masses. The color content of the tumor scan (color score) at color Doppler ultrasound examination was higher in the malignant masses. Use of hormonal therapy was less common among women with a malignant than a benign unclassifiable mass, and CA 125 values were higher.

The only variable that was retained in a multivariate logistic regression model to calculate the risk of

Table 2 Clinical and ultrasound information for benign and malignant unclassifiable (difficult) masses

Variables	*Benign* (n = *171)*	*Malignant* (n = *73)*	P*
Clinical variables			
Age (years)	51 ± 16	52 ± 15	0.76
Parity	2 (0–9)	1 (0–5)	0.41
Postmenopausal	96 (56)	40 (55)	0.85
Hysterectomy	24 (14)	4 (5)	0.08
Hormonal replacement therapy	29 (17)	4 (5)	0.01
Personal history of ovarian cancer	4 (2)	5 (7)	0.13
Family history of ovarian cancer	3 (2)	2 (3)	0.64
Personal history of breast cancer	7 (4)	3 (4)	1
Family history of breast cancer	22 (13)	6 (8)	0.38
CA 125 (U/mL)†	18 (3–40 140)	39 (3–9814)	< 0.001
Gray-scale ultrasound variables			
Volume of lesion (mL)	129 (1.4–24 740)	215 (0.1–5269)	0.06
Bilateral	24 (14)	13 (18)	0.46
Ascites	10 (6)	6 (8)	0.57
Type of mass			0.18
Unilocular	9 (5)	1 (1)	
Unilocular solid	41 (24)	19 (26)	
Multilocular	26 (15)	7 (10)	
Multilocular solid	69 (40)	27 (37)	
Solid	26 (15)	19 (26)	
Multilocular with > 10 locules	8 (5)	4 (5)	0.76
Number of locules	2 (0 to > 10)	1 (0 to > 10)	0.16
Pain at ultrasound examination	31 (18)	11 (15)	0.56
Echogenicity of cyst fluid			0.17
Anechoic	47 (27)	12 (16)	
Low level	53 (31)	26 (36)	
Ground glass	19 (11)	9 (12)	
Hemorrhagic	2 (1)	—	
Mixed	24 (14)	7 (10)	
No cyst fluid	26 (15)	19 (26)	
Papillary projections present	82 (48)	34 (47)	0.84
Irregular papillation	53 (65)	22 (65)	0.99
Flow in papillation	33 (40)	20 (59)	0.07
Number of papillations	1 (1 to ≥ 4)	3 (1 to ≥ 4)	0.009
Height of papillation (mm)	9 (3–48)	11 (3–52)	0.13
Mass with solid components	136 (80)	65 (89)	0.06
Largest diameter of largest solid component (mm)	24 (4–230)	42 (4–270)	< 0.001
Volume of largest solid component (mL)	4 (0.02–1343)	20 (0.008–5269)	< 0.001
Ratio volume of largest solid component/volume lesion	0.03 (0.00–1)	0.11 (0.00–1)	0.007
Incomplete septum	10 (6)	4 (5)	1
Irregular walls	108 (63)	50 (68)	0.42
Shadows	16 (9)	6 (8)	0.77
Doppler ultrasound variables			
Color score			0.003
Color score 1	31 (18)	8 (11)	
Color score 2	72 (42)	22 (30)	
Color score 3	58 (34)	32 (44)	
Color score 4	10 (6)	11 (15)	
Venous flow only	20 (12)	6 (8)	0.41
Pulsatility index‡	0.80 (0.26–5.80)	0.78 (0.36–1.94)	0.53
Resistance index‡	0.54 (0.09–1)	0.53 (0.22–0.86)	0.76
Peak systolic velocity (cm/s)‡	11.9 (0.4–77)	15.3 (5.0–64)	0.13
Time-averaged maximum velocity (cm/s)‡	7.8 (0.56–60)	9.0 (3.0–45)	0.08

Continuous variables are given as mean ± SD or as median (range); all other results are given as *n* (%). *Student's *t*-test used for analysis of data on patient's age, Mann–Whitney *U*-test used for analyses of data for the other continuous variables and Fisher's exact test or chi-square test used for discrete variables. †CA 125 was measured in 191 patients: in 127 (74%) patients with a benign mass and in 64 (88%) patients with a malignant mass. ‡Pulsatility index, resistance index, peak systolic velocity and time-averaged maximum velocity could be measured in 179 patients: in 120 (70%) patients with a classifiable mass and in 59 (81%) patients with an unclassifiable mass.

malignancy in an unclassifiable adnexal mass was the largest diameter (in mm) of the largest solid component: OR = 1.04; 95% CI, 1.02–1.06. This means that for every mm increase in the largest diameter of the largest solid component, the odds of malignancy increased by 4%. The probability of malignancy was equal to $y = 1/(1 + e^{-z})$, where $z = -1.950 + 0.037 \times$ (the largest diameter of the largest solid component). The ROC curve of the

Table 3 Diagnostic performance of subjective assessment, a logistic regression model, the risk of malignancy index (RMI) and CA 125 in unclassifiable masses

	All unclassifiable masses		*Unclassifiable masses with available CA 125*
Variable	*Training set* (n = *160)*	*Test set* (n = *84)*	*Test set* (n = *68)*
Subjective assessment			
Sensitivity	71 (32/45) (56–84)	68 (19/28) (48–84)	65 (17/26) (44–83)
Specificity	60 (69/115) (50–69)	59 (33/56) (45–72)	55 (23/42) (39–70)
LR+	1.78 (1.33–2.38)	1.65 (1.10–2.48)	1.45 (0.94–2.23)
LR−	0.48 (0.30–0.78)	0.55 (0.31–0.98)	0.63 (0.35–1.15)
Logistic regression model (cut-off 0.25*)			
AUC	0.68 (0.59–0.78)	0.65 (0.53–0.78)	0.61 (0.47–0.76)
Sensitivity	71 (32/45) (56–84)	64 (18/28) (44–81)	62 (16/26) (41–80)
Specificity	57 (66/115) (48–67)	55 (31/56) (41–69)	52 (22/42) (36–68)
LR+	1.67 (1.26–2.21)	1.44 (0.96–2.15)	1.29 (0.83–2.01)
LR−	0.50 (0.31–0.82)	0.65 (0.37–1.12)	0.73 (0.42–1.29)
Logistic regression model (cut-off 0.36†)			
Sensitivity	60 (27/45) (44–74)	50 (14/28) (31–69)	54 (14/26) (33–73)
Specificity	77 (88/115) (68–84)	73 (41/56) (60–84)	71 (30/42) (55–84)
LR+	2.56 (1.70–3.84)	1.87 (1.06–3.30)	1.88 (1.04–3.42)
LR−	0.52 (0.36–0.76)	0.68 (0.46–1.02)	0.65 (0.41–1.02)
RMI (cut-off 200)			
AUC			0.54 (0.39–0.69)
Sensitivity			35 (9/26) (17–56)
Specificity			71 (30/42) (55–84)
LR+			1.21 (0.59–2.47)
LR−			0.92 (0.65–1.28)
CA 125 (cut-off 35 U/mL)			
AUC			0.61 (0.47–0.76)
Sensitivity			50 (13/26) (30–70)
Specificity			60 (25/42) (43–74)
LR+			1.24 (0.73–2.10)
LR−			0.84 (0.53–1.33)

Results are given as % (*n*/*n*) (95% CI) for sensitivity and specificity values and as value (95% CI) for area under the receiver–operating characteristics curve (AUC), positive likelihood ratio (LR+) and negative likelihood ratio (LR−). *The risk cut-off that yielded the same sensitivity as subjective assessment in the training set (i.e. 71%). This risk cut-off corresponds to the largest diameter of the largest solid component of 23 mm. †The mathematically 'optimal' risk cut-off according to the shape of the receiver–operating characteristics curve. This risk cut-off corresponds to the largest diameter of the largest solid component of 37 mm.

logistic regression model, when applied to the training set and test set, is shown in Figure 4. The AUC was 0.68 (95% CI, 0.59–0.78) on the training set and 0.65 (95% CI, 0.53–0.78) on the test set. When we restricted our logistic regression analysis to the 123 masses in the training set with available CA 125 results, the variables 'largest diameter of the largest solid component' and log(CA 125) were retained in the model. However the AUC of the model containing these two variables was not larger than that of the model containing only the variable 'largest diameter of the largest solid component' (AUC 0.73 vs. 0.69, $P = 0.12$).

In unclassifiable masses ($n = 244$), the sensitivity with regard to malignancy of subjective assessment was 70% (51/73), the specificity was 60% (102/171), the LR+ was 1.73 and the LR− was 0.51. For the classifiable masses ($n = 3267$), the corresponding values were 91% (802/878), 96% (2292/2389), 22.50 and 0.09 ($P < 0.001$ for the difference in sensitivity and $P < 0.001$ for the difference in specificity). The diagnostic performance of subjective assessment, the logistic regression model, CA 125 and the RMI when tested on the training set and the test set are shown in Table 3. No test could discriminate with reasonable accuracy[21] between benign and malignant unclassifiable masses.

DISCUSSION

We have shown that about 7% of adnexal masses currently considered appropriate for surgical removal cannot be classified by an experienced ultrasound examiner as benign or malignant using subjective assessment of gray-scale and Doppler ultrasound findings (pattern recognition). Multilocular cysts with more than 10 locules and masses with small solid components seem to be more difficult to classify than other types of tumor. The histological diagnoses that presented the greatest diagnostic difficulties were borderline tumors, cystadeno(fibro)mas and fibromas. Even though we succeeded in creating a logistic regression model to predict malignancy in masses that could not be confidently classified as benign or malignant by using subjective assessment, our model had very poor diagnostic performance. It was not superior to subjective assessment. CA 125 and RMI also performed poorly.

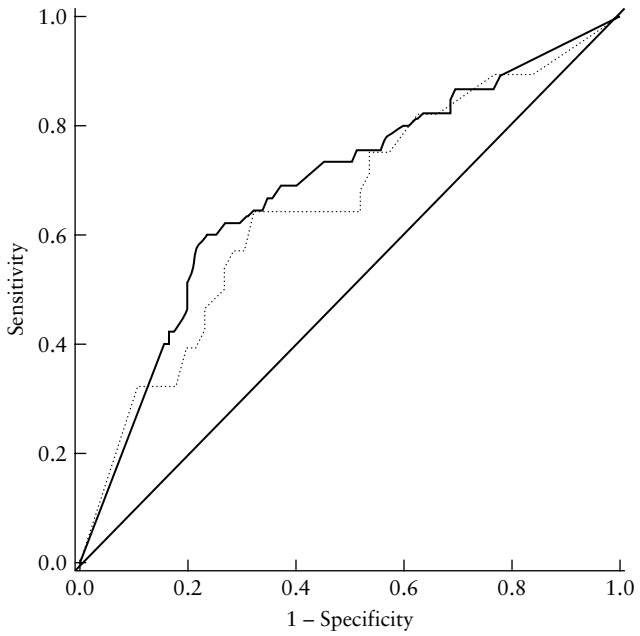

Figure 4 Receiver–operating characteristics curves for the logistic regression model on the training set (——, $n = 160$) and the test set (·····, $n = 84$) of unclassifiable masses. The area under the curve is 0.68 (95% CI, 0.59–0.78) on the training set and 0.65 (95% CI, 0.53–0.78) on the test set. The probability of malignancy is equal to $y = 1/(1 + e^{-z})$, where $z = -1.950 + 0.037 \times$ (the largest diameter of the largest solid component).

It is a strength of our study that it was large and involved many centers. Therefore, our results are likely to be generalizable to other experienced ultrasound examiners using good ultrasound systems, provided that they are exposed to a population similar to our study population. We believe that our sample of masses is representative of the types of extrauterine pelvic mass that it is currently considered appropriate to remove surgically. However, even though our study was large and multicenter, the number of unclassifiable masses was rather small. This is explained by few adnexal masses being difficult to classify as benign or malignant using subjective assessment. It is also a limitation that the diagnostic performance of subjective assessment, CA 125 and RMI were tested prospectively, while the performance of the logistic regression model (i.e. the performance of the largest diameter of the largest solid component of the tumor) was not truly prospectively evaluated even though it was tested on a test set separate from the training set. Therefore, a comparison of our logistic regression model with the other methods is not a fair comparison because the diagnostic performance of our logistic regression model may have been overestimated relative to the other methods. This, however, does not invalidate our conclusion (that all diagnostic tests performed very poorly and that none was clinically useful). Another weakness of our study was that the CA 125 level was not measured in all women with tumors and was measured more often in those with malignant tumors. Even though this may have introduced bias, it does not invalidate our results. When the diagnostic performance of subjective assessment, the logistic regression model, CA 125 and RMI was tested on the test set of tumors for which CA 125 results were available, none of the methods performed well. It seems unlikely that CA 125 and/or RMI would have performed very differently among the 44 benign tumors and nine malignant tumors for which CA 125 values were missing (there was a higher proportion of serous cystadenomas but a lower proportion of borderline tumors among tumors with missing than available CA 125 results). Another limitation of our study was that different CA 125 kits were used to assess the level of serum CA 125. However, this reflects clinical reality, and there is some evidence that the variation in CA 125 resulting from the use of different kits is not large[22,23].

Our results confirm those of our previous smaller study[7]. In the smaller study, 8% of all masses were unclassifiable, and borderline tumors and (papillary) cystadenomas/cystadenofibromas were over-represented among the unclassifiable masses. In the smaller study, no logistic regression model could be constructed to discriminate between benign and malignant unclassifiable masses. This is likely to be explained by the small number of cases.

Our logistic regression model showed that the larger the solid components in an unclassifiable adnexal mass, the greater the risk of malignancy. However, this finding has very limited clinical importance because the model performed poorly (see also the substantial overlap in the size of the largest solid component between benign and malignant unclassifiable masses in Table 2). Our findings are disappointing because if a mass cannot be reliably classified as benign it is likely to be treated as potentially malignant. This means that many women with an unclassifiable mass are likely to undergo unnecessarily extensive surgical procedures with their associated complications. It is less likely that an uncertain diagnosis will result in initial 'undertreatment', necessitating secondary, more extensive, surgery, but this could also happen in some cases. It is disappointing that neither RMI nor CA 125 was helpful in discriminating between benign and malignant masses that could not be confidently classified as benign or malignant by subjective assessment. It remains to be seen if there are other methods that can work as secondary diagnostic tests in the group of unclassifiable adnexal masses, for example new tumor markers possibly developed by proteomics[24], three-dimensional ultrasound with characterization of the vessel tree of tumors[25] or examination with ultrasound contrast[26].

ACKNOWLEDGMENTS

This research was supported by the Research Council of the Katholieke Universiteit Leuven: GOA MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC), PFV/10/002 (OPTEC); the Flemish Government: FWO G.0302.07 (SVM), IWT-TBM070706-IOTA3; the Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011); the Swedish Medical Research Council (grant 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; and two Swedish governmental grants (ALF-medel and Landstingsfinansierad Regional Forskning).

APPENDIX

IOTA Steering Committee

Dirk Timmerman, *Leuven, Belgium*
Lil Valentin, *Malmö, Sweden*
Tom Bourne, *London, UK*
Antonia C. Testa, *Rome, Italy*
Sabine Van Huffel, *Leuven, Belgium*
Ignace Vergote, *Leuven, Belgium*

Principal investigators and recruitment centers

Jean-Pierre Bernard, *Centre medical des Pyramides, Maurepas, France*;
Nicoletta Colombo, *European Institute of Oncology, Milan, Italy*;
Artur Czekierdowski, *Medical University in Lublin, Lublin, Poland*;
Elisabeth Epstein, *Lund University Hospital, Lund University, Lund, Sweden*;
Daniela Fischerova, *General Faculty Hospital of Charles University, Prague, Czech Republic*;
Robert Fruscio, *Ospedale S. Gerardo, Universita di Milano Bicocca, Monza, Italy*;
Stefano Greggi, *Instituto Nationale dei Tumori, Fondazione Pascale, Naples, Italy*;
Stefano Guerriero, *University of Cagliari, Ospedale San Giovanni di Dio, Cagliari, Italy*;
Jingzhang, *Chinese PLA General Hospital, Beijing, China*;
Davor Jurkovic, *King's College Hospital, London, UK*;
Fabrice Lécuru, *Hopital Europeen Georges Pompidou, Paris, France*;
Francesco Leone, *Clinical Sciences Institute L. Sacco, University of Milan, Milan, Italy*;
Andrea Alberto Lissoni, *Ospedale S. Gerardo, Universita di Milano Bicocca, Monza, Italy*;
Ulrike Metzger, *Hopital Europeen Georges Pompidou, Paris, France*;
Henry Muggah, *McMaster University, St Joseph's Hospital, Hamilton, Ontario, Canada*;
Dario Paladini, *Universita degli Studi di Napoli, Naples, Italy*;
Alberto Rossi, *Universita degli Studi di Udine, Udine, Italy*;
Luca Savelli, *University of Bologna, Bologna, Italy*;
Antonia Testa, *Universita Cattolica del Sacro Cuore, Rome, Italy*;
Dirk Timmerman, *University Hospitals Leuven, Leuven, Belgium*;
Diego Trio, *Macedonio Melloni Hospital, University of Milan, Milan, Italy*;
Lil Valentin, *Skane University Hospital Malmö, Lund University, Malmö, Sweden*;
Caroline Van Holsbeke, *Ziekenhuis Oost-Limburg, Genk, Belgium*.

REFERENCES

1. Medeiros LR, Stein AT, Fachel J, Garry R, Furness S. Laparoscopy versus laparotomy for benign ovarian tumours. *Cochrane Database Syst Rev* 2005; **20**: CD004751.
2. Carley ME, Klingele CJ, Gebhart JB, Webb MJ, Wilson TO. Laparoscopy versus laparotomy in the management of benign unilateral adnexal masses. *J Am Assoc Gynecol Laparosc* 2002; **9**: 321–326.
3. Cannistra S. Ovarian cancer. *N Engl J Med* 2004; **351**: 2519–2529.
4. Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P, Gore ME, Kaern J, Verrelst H, Sjövall 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.
5. 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.
6. 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.
7. Valentin L, Ameye L, Jurkovic D, Metzger U, Lécuru F, Sabine 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.
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. Van Holsbeke C, Van Calster B, Valentin L, Testa AC, Ferrazzi E, Dimou I, Lu C, Moerman P, Van Huffel S, Vergote I, Timmerman D; International Ovarian Tumor Analysis Group. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the International Ovarian Tumor Analysis Group. *Clin Cancer Res* 2007; **13**: 4440–4447.
10. 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.
11. 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.
12. Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, Van Holsbeke C, Fruscio R, Czekierdowski A, Jurkovic D, Savelli L, Vergote I, Bourne T, Van Huffel S, Valentin L. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound Obstet Gynecol* 2010; **36**: 226–234.
13. 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.
14. Kenemans P, van Kamp GJ, Oehr P, Verstraeten RA. Heterologous double-determinant immunoradiometric assay CA 125

II: reliable second-generation immunoassay for determining CA 125 in serum. *Clin Chem* 1993; **39**: 2509–2513.
15. Shepherd JH. Revised FIGO staging for gynaecological cancer. *Br J Obstet Gynaecol* 1989; **96**: 889–892.
16. Heintz AP, Odicino F, Maisonneuve P, Beller U, Benedet JL, Creasman WT, Ngan HY, Pecorelli S. Carcinoma of the ovary. *Int J Gynaecol Obstet* 2003; **83** (Suppl. 1): 135–166.
17. Westfall PH, Wolfinger RD. Multiple tests with discrete distributions. *Am Stat* 1997; **51**: 3–8.
18. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol* 1991; **44**: 763–770.
19. Duffy MJ, Bonfrer JM, Kulpa J, Rustin GJ, Soletormos G, Torre GC, Tuxen MK, Zwirner M. CA125 in ovarian cancer: European Group on Tumor Markers guidelines for clinical use. *Int J Gynecol Cancer* 2005; **15**: 679–691.
20. Delong ER, Delong DM, Clarkepearson DI. Comparing the areas under 2 or more correlated receiver operating characteristic curves – a nonparametric approach. *Biometrics* 1988; **44**: 837–845.
21. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA* 1994; **271**: 703–707.
22. Bonfrer J, Baan A, Jansen E, Lentfer D, Kenemans P. Technical evaluation of three second generation CA 125 assays. *Eur J Clin Chem Clin Biochem* 1994; **32**: 201–207.
23. Davelaar E, van Kamp G, Verstraeten R, Kenemans P. Comparison of seven immunoassays for the quantification of CA 125 antigen in serum. *Clin Chem* 1998; **44**: 1417–1422.
24. Cadron I, Van Gorp T, Timmerman D, Amant F, Waelkens E, Vergote I. Application of proteomics in ovarian cancer: which sample should be used? *Gynecol Oncol* 2009; **115**: 497–503.
25. Sladkevicius P, Jokubkiene L, Valentin L. Contribution of morphological assessment of the vessel tree by three-dimensional ultrasound to a correct diagnosis of malignancy in ovarian masses. *Ultrasound Obstet Gynecol* 2007; **30**: 874–882.
26. Testa AC, Timmerman D, Van Belle V, Fruscella E, Van Holsbeke C, Savelli L, Ferrazzi E, Leone FP, Marret H, Tranquart F, Exacoustos C, Nazzaro G, Bokor D, Magri F, Van Huffel S, Ferrandina G, Valentin L. Intravenous contrast ultrasound examination using contrast-tuned imaging (CnTI) and the contrast medium SonoVue for discrimination between benign and malignant adnexal masses with solid components. *Ultrasound Obstet Gynecol* 2009; **34**: 699–710.

SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:

Table S1 Clinical and ultrasound information for classifiable and unclassifiable masses

