Ultrasound Obstet Gynecol 2010; **36**: 226–234
Published online 19 July 2010 in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/uog.7636

Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group

D. TIMMERMAN*, B. VAN CALSTER†, A. C. TESTA‡, S. GUERRIERO§, D. FISCHEROVA¶, A. A. LISSONI**, C. VAN HOLSBEKE*††, R. FRUSCIO**, A. CZEKIERDOWSKI‡‡, D. JURKOVIC§§, L. SAVELLI¶¶, I. VERGOTE*, T. BOURNE*#, S. VAN HUFFEL† and L. VALENTIN***

**Department of Obstetrics and Gynecology, University Hospitals and †Department of Electrical Engineering, ESAT-SISTA, Katholieke Universiteit Leuven, Leuven and ††Department of Obstetrics and Gynecology, Ziekenhuis Oost-Limburg, Genk (ZOL), Genk, Belgium, ‡Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Roma, §Department of Obstetrics and Gynecology, University of Cagliari, Sardinia, **Clinica Ostetrica e Ginecologica, Ospedale S. Gerardo, Università di Milano Bicocca, Monza and ¶¶Reproductive Medicine Unit, Department of Obstetrics and Gynecology, University of Bologna, Bologna, Italy, ¶Oncogynecological Center, Department of Obstetrics and Gynecology, General Faculty Hospital of Charles University, Prague, Czech Republic, ‡‡1st Department of Gynecologic Oncology and Gynecology, Medical University in Lublin, Poland, §§Department of Obstetrics and Gynaecology, University College Hospital and #Department of Obstetrics and Gynaecology, Imperial College London, Hammersmith Campus, London, UK and ***Department of Obstetrics and Gynecology, Malmö University Hospital, Lund University, Malmö, Sweden*

KEYWORDS: color Doppler ultrasonography; logistic models; ovarian neoplasms; sensitivity and specificity; ultrasonography

ABSTRACT

Objectives *The aims of the study were to temporally and externally validate the diagnostic performance of two logistic regression models containing clinical and ultrasound variables in order to estimate the risk of malignancy in adnexal masses, and to compare the results with the subjective interpretation of ultrasound findings carried out by an experienced ultrasound examiner ('subjective assessment').*

Methods *Patients with adnexal masses, who were put forward by the 19 centers participating in the study, underwent a standardized transvaginal ultrasound examination by a gynecologist or a radiologist specialized in ultrasonography. The examiner prospectively collected information on clinical and ultrasound variables, and classified each mass as benign or malignant on the basis of subjective evaluation of ultrasound findings. The gold standard was the histology of the mass with local clinicians deciding whether to operate on the basis of ultrasound results and the clinical picture. The models' ability to discriminate between malignant and benign masses was assessed, together with the accuracy of the risk estimates.*

Results *Of the 1938 patients included in the study, 1396 had benign, 373 had primary invasive, 111 had borderline malignant and 58 had metastatic tumors. On external validation (997 patients from 12 centers), the area under the receiver–operating characteristics curve (AUC) for a model containing 12 predictors (LR1) was 0.956, for a reduced model with six predictors (LR2) was 0.949 and for subjective assessment was 0.949. Subjective assessment gave a positive likelihood ratio of 11.0 and a negative likelihood ratio of 0.14. The corresponding likelihood ratios for a previously derived probability threshold (0.1) were 6.84 and 0.09 for LR1, and 6.36 and 0.10 for LR2. On temporal validation (941 patients from seven centers), the AUCs were 0.945 (LR1), 0.918 (LR2) and 0.959 (subjective assessment).*

Conclusions *Both models provide excellent discrimination between benign and malignant masses. Because the models provide an objective and reasonably accurate risk estimation, they may improve the management of women with suspected ovarian pathology. Copyright © 2010 ISUOG. Published by John Wiley & Sons, Ltd.*

INTRODUCTION

Every year in the USA an estimated 289 000 women undergo surgery for an ovarian cyst or pelvic mass[1]. For the optimal management of these patients it is very important to distinguish malignant from benign tumors. The implications of a misdiagnosis are significant. Patients thought to have a malignant ovarian tumor may undergo extensive staging procedures[2], which are

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: 3 March 2010

expensive and distressing for the patient. Performing laparoscopic surgery in early stage ovarian cancer should be avoided because rupture during surgery on a Stage I ovarian cancer may worsen the prognosis[3]. Most benign cysts can be treated with minimally invasive surgery, which is associated with a shorter duration of hospitalization and rehabilitation than laparotomy[4,5].

Both the morphology and vascularity of an ovarian mass described using ultrasound examination can be used to predict the likelihood of malignancy[6–9]. The Risk of Malignancy Index[10] was the first prediction model to combine clinical, ultrasound and tumor marker information. Many more prediction models to estimate the risk of malignancy, including logistic regression models, have since been developed. The diagnostic performance of these algorithms was reasonable when evaluated prospectively in different centers,[11] but their performance was poorer than that of subjective interpretation of ultrasound findings ('subjective assessment') by experienced ultrasound examiners[12–14].

The multicenter International Ovarian Tumor Analysis (IOTA) study was designed to create improved risk-prediction models to discriminate between benign and malignant adnexal tumors. In the first study, of 1066 patients with adnexal tumors from nine centers, we developed two logistic regression models, LR1 and LR2, to estimate the likelihood of malignancy (IOTA Phase 1 study)[15]. The models, which consist of 12 and six variables, respectively, performed very well in the test group of that study (internal validation)[15] and in a small temporal validation study using three centers[13], but their performance has not been externally validated prospectively. This is an essential step before such models can be used in clinical practice[16–19].

The primary aim of the present study was to prospectively validate the diagnostic performance of the logistic regression models LR1 and LR2 for discriminating between malignant and benign adnexal tumors, both temporally, using centers that were also involved in model development, and externally, using unrelated clinical centers. The secondary aim was to compare the diagnostic performance of these models with that of subjective assessment carried out by experienced ultrasound examiners.

METHODS

In this prospective multicenter study, the IOTA Phase 2 study, we examined the ability of two logistic regression models (LR1 and LR2) and subjective interpretation of ultrasound findings carried out by an expert ('subjective assessment') to discriminate between benign and malignant adnexal masses on a population of women who were operated on for an adnexal mass. Patients with an adnexal mass underwent a standardized gray scale and Doppler ultrasound examination by a gynecologist or radiologist specialized in gynecological ultrasound examinations (i.e. the principal investigator at each center), as described below. Only patients who were operated on ≤ 120 days after the ultrasound examination were included, the outcome variable being the histological diagnosis of the mass. The decision of whether or not to operate was made by local clinicians on the basis of the results of the ultrasound examination, the clinical picture and local management protocols.

Nineteen centers in eight countries participated in the study (the centers are listed in the Appendix). Seven of these centers also contributed patients for the development of LR1 and LR2. The data from these seven centers were used for temporal validation. The other 12 centers were used for external validation.

Our research protocol was ratified by the local Ethics Committee at each recruitment center, and we followed the guidelines of the STARD (Standards for Reporting of Diagnostic Accuracy) initiative when writing this paper[20].

Patients

Inclusion criteria

Inclusion criteria were patients presenting with at least one adnexal mass who underwent an ultrasound examination by a principal investigator at one of the participating centers. 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 mass, or the one most easily accessible by ultrasound examination, was included.

Exclusion criteria

Exclusion criteria were: pregnancy; refusal of transvaginal ultrasonography; and failure to undergo surgical removal of the mass within 120 days of the ultrasound examination.

Data collection

A dedicated, secure electronic data-collection system was developed for the study (IOTA 2 Study Screen; Astraia Software, Munich, Germany). A unique identifier was generated automatically for each patient's record. Clinicians at each center could only view or update patient records from their own center. Data security was ensured by not transferring the patient's name and by encrypting all data communication. Data integrity and completeness were ensured by client-side checks in the system supplied by Astraia Software and manual checks by one biostatistician and two expert ultrasound examiners.

Clinical variables

Before the ultrasound examination a standardized history was taken by the ultrasound examiner in the same manner as in the IOTA Phase 1 study[15]. It included information on the patient's history of ovarian and breast cancer, the number of first-degree relatives with ovarian or breast

cancer, the patient's age, menopausal status, day of menstrual cycle (if appropriate), current hormone therapy and previous gynecological surgery. Women ≥ 50 years of age, who had undergone hysterectomy, were defined as postmenopausal.

Ultrasound examination

In all cases a transvaginal ultrasound scan was performed in the same standardized manner as in the IOTA Phase 1 study[15]. Transabdominal ultrasonography was used to examine a large mass that could not be seen in its entirety using a transvaginal probe. Gray scale and color Doppler ultrasound imaging was used to obtain information on more than 40 morphological and blood-flow variables to characterize each adnexal mass. The presence or absence of pain during the examination was recorded. Details on the ultrasound examination technique and the ultrasound terms and definitions used have been described elsewhere[21]. Finally, the ultrasound examiner, on the basis of subjective interpretation of the ultrasound findings, first stated whether the mass was likely to be malignant or benign and then, as a second step, whether the suggested diagnosis was certain, probable or uncertain. This resulted in six levels of diagnostic confidence: certainly malignant, probably malignant, uncertain but considered malignant, uncertain but considered benign, probably benign, and certainly benign. CA 125 results were not available to the ultrasound examiner at the time of the ultrasound examination.

The ultrasound information was recorded prospectively in the electronic data-collection system and was locked at the time of the examination and so could not be changed thereafter. The risk calculation by the logistic regression models was performed centrally after the conclusion of the study, and so the logistic models had no role in the decision-making process.

Outcome measures

The final outcome measure of the study was the histological diagnosis. Surgery was performed by laparoscopy or laparotomy, according to the surgeon's judgment. The excised tissues underwent histological examination at the local center. Tumors were classified according to the criteria recommended by the International Federation of Gynecology and Obstetrics[22].

The logistic regression models

The original LR1[15] contains 12 variables: (1) age of the patient (years); (2) the presence of ascites (yes = 1, no = 0); (3) the presence of blood flow within a papillary projection (yes = 1, no = 0); (4) largest diameter of the solid component (expressed in mm but with no increase above 50 mm); (5) irregular internal cyst walls (yes = 1, no = 0); (6) the presence of acoustic shadows (yes = 1, no = 0); (7) personal history of ovarian cancer (yes = 1, no = 0); (8) current hormonal therapy (yes = 1, no = 0); (9) largest diameter of the lesion (mm); (10) the presence of pain during the examination (yes = 1, no = 0); (11) the presence of a purely solid tumor (yes = 1, no = 0); and (12) the color score (1, 2, 3 or 4). The model's estimated probability of malignancy for an adnexal tumor equals $1/(1+e^{-z})$, where $z = -6.7468 + 0.0326(1) + 1.5513(2) + 1.1737(3) + 0.0496(4) + 1.1421(5) - 2.3550(6) + 1.5985(7) - 0.9983(8) + 0.00841(9) - 0.8577(10) + 0.9281(11) + 0.4916(12)$, and e is the mathematical constant and base value of natural logarithms.

A simpler model (LR2) uses only those six variables that were first entered into the model LR1 when using stepwise selection of variables. The variables included in LR2 are: (1) age of the patient (years); (2) the presence of ascites (yes = 1, no = 0); (3) the presence of blood flow within a papillary projection (yes = 1, no = 0); (4) maximal diameter of the solid component (expressed in mm but with no increase above 50 mm); (5) irregular internal cyst walls (yes = 1, no = 0); and (6) the presence of acoustic shadows (yes = 1, no = 0). For LR2, $z = -5.3718 + 0.0354(1) + 1.6159(2) + 1.1768(3) + 0.0697(4) + 0.9586(5) - 2.9486(6)$.

Statistical analysis

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

Performance of the models was assessed using measures of discrimination and calibration (i.e. the correctness of the predicted probabilities of malignancy)[23]. Subjective assessment was evaluated in relation to discriminatory ability only. Discriminatory performance was evaluated using receiver–operating characteristics (ROC) curves and the area under the curve (AUC). The six levels of diagnostic confidence were used to construct the ROC curve for subjective assessment. Using a risk cut-off of 0.1 to predict malignancy, the sensitivity, specificity, positive likelihood ratio (LR+) and negative likelihood ratio (LR−) for LR1 and LR2 were computed. This risk cut-off was the one suggested for clinical use in the original publication where the development of LR1 and LR2 was described[15]. For subjective assessment, these measures were computed using the ultrasound examiner's dichotomous classification of the tumor as benign or malignant.

Calibration[23] of the predicted probabilities of malignancy calculated by the models was investigated using calibration curves and by the ratio of average predicted risk (i.e. the average probability of malignancy) to the observed risk (i.e. the prevalence of malignancy)[24]. A ratio of < 1 suggests general underprediction of the risk, whereas a ratio of > 1 suggests general overprediction of the risk. The calibration curves link the predicted risk with the observed outcome using a non-parametric logistic regression model based on local regression[25].

RESULTS

Between November 2005 and October 2007, 1970 patients fulfilled our inclusion criteria. Of these, 32

(1.6%) were excluded for the following reasons: no surgical removal of the mass within 120 days after the ultrasound examination ($n = 15$); pregnant at the time of the examination ($n = 12$); errors in data entry ($n = 4$); and protocol violation ($n = 1$). Thus, 1938 patients were included: 1396 (72%) had benign tumors, 111 (5.7%) had borderline malignant tumors, 373 (19.2%) had primary invasive tumors and 58 (3%) had secondary tumors of the ovary (i.e. metastatic invasive tumors). The number of patients and histological outcomes per center are presented in Table 1. There were 941 patients for the temporal validation with a prevalence of malignancy of 30%, and 997 patients for the external validation with a prevalence of malignancy of 26%. Details on histological diagnoses are shown in Table 2, and demographic background information, clinical and ultrasound information are shown in Table 3.

External validation

The diagnostic performance of LR1 and LR2 and their subjective assessment in the 12 new centers ($n = 997$) is shown in Table 4 and Figure 1. The AUCs of the three methods were 0.956 (95% CI, 0.940–0.968), 0.949 (95% CI, 0.931–0.964) and 0.949 (95% CI, 0.930–0.964), respectively. The AUCs were slightly smaller in postmenopausal patients than in premenopausal patients (Table S1). The AUCs of LR1 and LR2 in each center are shown in Table S2. In all but one center the AUC was larger than 0.91.

Using a risk cut-off of 0.10 (derived from our original report) for LR1 to predict malignancy[15], LR1 missed 20 cancers (seven borderline tumors, three metastatic tumors in the ovary and 10 primary invasive ovarian malignancies (of which one was a Stage I epithelial cancer, five were Stage III epithelial cancers and four

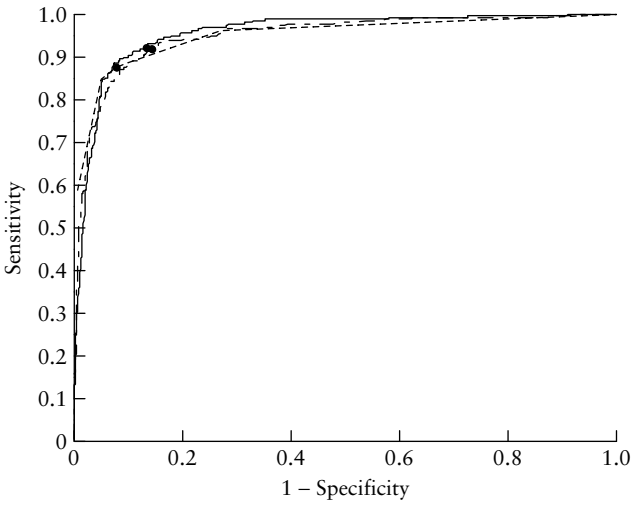

Figure 1 Receiver–operating characteristics (ROC) curves for logistic regression models LR1 (——) and LR2 (— - —) and for subjective assessment using pattern recognition (- - - - -) on external validation ($n = 997$). The dots in the ROC curves for LR1 and LR2 correspond to the sensitivity and false-positive rate (1 minus specificity) for the thresholds (i.e. an estimated risk of malignancy of 0.10) recommended in the study in which the models were created[12]. The ROC curve for subjective assessment was created using six levels of diagnostic confidence. The dot in the ROC curve for subjective assessment corresponds to sensitivity and false-positive rate of the ultrasound examiner's dichotomous classification of the tumor as benign or malignant. Areas under the curves were 0.956 for LR1, 0.949 for LR2 and 0.949 for pattern recognition.

Table 1 Histological diagnosis according to participating center

Center	*Diagnosis* (n *(%))*				*Total* (n *(%))*
	Benign	*Primary invasive*	*Borderline*	*Metastatic invasive*	
External validation (new centers)	742 (74.4)	187 (18.8)	42 (4.2)	26 (2.6)	997 (100)
Genk, Belgium	173 (87)	21	5	1	200
Lublin, Poland	101 (66)	45	3	5	154
Cagliari, Italy	134 (87)	13	3	4	154
Bologna, Italy	124 (92)	6	3	2	135
Milan (B), Italy	41 (44)	36	10	7	94
Prague, Czech Republic	39 (43)	35	15	1	90
Beijing, China	57 (78)	12	1	3	73
Lund, Sweden	31 (82)	4	1	2	38
Milan (C), Italy	17 (81)	3	1	0	21
Udine, Italy	10 (59)	6	0	1	17
Ontario, Canada	11 (92)	1	0	0	12
Naples (B), Italy	4 (44)	5	0	0	9
Temporal validation (old centers)	654 (69.5)	186 (19.8)	69 (7.3)	32 (3.4)	941 (100)
Leuven, Belgium	155 (62)	60	24	13	252
Monza, Italy	199 (79)	31	17	4	251
Malmö, Sweden	110 (80)	21	6	0	137
Rome, Italy	54 (44)	49	11	8	122
London, UK	40 (62)	13	9	3	65
Naples (A), Italy	51 (80)	8	2	3	64
Milan (A), Italy	45 (90)	4	0	1	50
Total	1396 (72.0)	373 (19.2)	111 (5.7)	58 (3.0)	1938 (100)

Percentages are calculated per row.

Table 2 Details on histological diagnoses

Histological diagnosis	*All* (n = *1938*) (n *(%)*)	*Menopausal status* (n *(%)*)		*Center* (n *(%)*)	
		Premenopausal (n = *1197*)	*Postmenopausal* (n = *741*)	*Old centers* (n = *941*)	*New centers* (n = *997*)
Benign	1396 (72.0)	1014 (84.8)	382 (51.5)	654 (69.5)	742 (74.4)
Endometrioma	400 (20.6)	382 (31.9)	18 (2.4)	192 (20.4)	208 (20.9)
Serous cystadenoma	236 (12.2)	103 (8.6)	133 (17.9)	126 (13.4)	110 (11.0)
Teratoma	226 (11.7)	195 (16.3)	31 (4.2)	96 (10.2)	130 (13.0)
Mucinous cystadenoma	138 (7.1)	70 (5.9)	68 (9.2)	68 (7.2)	70 (7.0)
Simple cyst + parasalpingeal cyst	131 (6.8)	85 (7.1)	46 (6.2)	56 (6.0)	75 (7.5)
Fibroma	86 (4.4)	30 (2.5)	56 (7.5)	44 (4.7)	42 (4.2)
Functional cyst	77 (4.0)	68 (5.7)	9 (1.2)	27 (2.9)	50 (5.0)
Hydrosalpinx + salpingitis	49 (2.5)	47 (3.9)	2 (0.3)	22 (2.3)	27 (2.7)
Abscess	24 (1.2)	17 (1.4)	7 (0.9)	8 (0.9)	16 (1.6)
Rare benign	18 (0.9)	8 (0.7)	10 (1.3)	11 (1.2)	7 (0.7)
Peritoneal pseudocyst	11 (0.6)	9 (0.8)	2 (0.3)	4 (0.4)	7 (0.7)
Malignant	542 (28.0)	182 (15.2)	360 (48.5)	287 (30.5)	255 (25.6)
Primary invasive	373 (19.2)	101 (8.4)	272 (36.7)	186 (19.8)	187 (18.8)
Stage I	70 (3.6)	23 (1.9)	47 (6.3)	32 (3.4)	38 (3.8)
Stage II	30 (1.6)	7 (0.6)	23 (3.1)	12 (1.3)	18 (1.8)
Stage III	202 (10.4)	41 (3.4)	161 (21.7)	105 (11.2)	97 (9.7)
Stage IV	30 (1.6)	8 (0.7)	22 (3.0)	14 (1.5)	16 (1.6)
Rare primary invasive	41 (2.1)	22 (1.8)	19 (2.6)	23 (2.4)	18 (1.8)
Borderline	111 (5.7)	62 (5.2)	49 (6.6)	69 (7.3)	42 (4.2)
Stage I	99 (5.1)	52 (4.3)	47 (6.3)	63 (6.7)	36 (3.6)
Stage II	3 (0.2)	2 (0.2)	1 (0.1)	1 (0.1)	2 (0.2)
Stage III	8 (0.4)	7 (0.6)	1 (0.1)	4 (0.4)	4 (0.4)
Stage IV	1 (0.1)	1 (0.1)	0	1 (0.1)	0
Metastatic	58 (3.0)	19 (1.6)	39 (5.3)	32 (3.4)	26 (2.6)

Table 3 Demographic, clinical and ultrasound characteristics of the study population

Variable	*External validation (new centers)*		*Temporal validation (old centers)*	
	Benign (n = *742*)	*Malignant* (n = *255*)	*Benign* (n = *654*)	*Malignant* (n = *287*)
Age (years, median)	40	56	41	57
Nulliparous	47.3	22.0	45.9	30.0
Personal history of ovarian cancer	0.4	3.5	1.2	4.5
Current use of hormonal therapy	12.0	6.3	13.5	7.3
Pain at ultrasound examination	24.1	11.0	14.5	11.5
Largest diameter of lesion (mm, median)	58	82	64	89
Solid component				
Present	27.2	94.9	29.8	89.9
Largest diameter (mm) if present (median)	24	51.5	26	54
Ascites	1.2	32.6	1.5	29.6
Papillations				
Present	10.0	31.4	11.8	38.0
Present, with blood flow	1.1	20.4	2.3	25.1
Irregular cyst walls	26.6	65.1	21.3	63.8
Acoustic shadows	19.1	5.9	15.3	4.2
Purely solid tumor	7.6	37.7	9.3	36.9
Subjective assessment				
Certainly benign	72.9	3.9	61.6	2.8
Probably benign	17.3	7.5	28.3	2.8
Uncertain, benign	1.9	1.2	3.5	1.4
Uncertain, malignant	3.0	2.8	2.9	8.0
Probably malignant	4.3	25.9	2.8	30.0
Certainly malignant	0.7	58.8	0.9	55.0

Data are expressed as % unless otherwise indicated.

were rare types of malignancy)) and yielded 100 false-positive results, resulting in an LR+ of 6.84 and an LR− of 0.09. LR2 missed 21 cancers (nine borderline tumors, two metastatic tumors in the ovary and 10 primary invasive ovarian malignancies (of which two were Stage I epithelial cancers, five were Stage III epithelial

Table 4 Results of prospective validation of the diagnostic performance of two logistic regression models used to calculate the risk of malignancy in adnexal masses (LR1 and LR2) and results of the subjective interpretation of ultrasound findings ('subjective assessment')

Validation	*Statistic*	*LR1*	*LR2*	*Difference LR1–LR2*	*Subjective assessment*	*Difference LR1– subjective assessment*	*Difference LR2– subjective assessment*
External	AUC (95% CI)	0.956 (0.940, 0.968)	0.949 (0.931, 0.964)	0.007 (−0.001, 0.015)	0.949 (0.930, 0.964)	0.007 (−0.010, 0.023)	0.000 (−0.019, 0.018)
(n = 997,	LR+ (95% CI)	6.84 (5.69, 8.25)	6.36 (5.33, 7.63)	0.48 (−0.36, 1.50)	11.0 (8.60, 14.1)	−4.2 (−7.5, −2.1)	−4.6 (−7.9, −2.4)
12 centers)	LR− (95% CI)	0.09 (0.06, 0.14)	0.10 (0.06, 0.14)	−0.01 (−0.03, 0.02)	0.14 (0.10, 0.19)	−0.05 (−0.09, 0.00)	−0.04 (−0.09, 0.01)
	Sensitivity (%)	92.2	91.8	0.4 (−2.1, 2.9)	87.5	4.7 (0.6, 9.1)	4.3 (0.0, 8.8)
	Invasive tumors only (%)	93.9	94.4	−0.5 (−3.3, 2.2)	89.2	4.7 (0.5, 9.2)	5.2 (0.7, 10.0)
	Specificity (%)	86.5	85.6	0.9 (−1.0, 2.9)	92.1	−5.6 (−8.0, −3.2)	−6.5 (−9.0, −4.0)
	Predicted vs. observed risk*	0.83	0.78				
Temporal	AUC (95% CI)	0.945 (0.930, 0.958)	0.918 (0.896, 0.936)	0.027 (0.017, 0.038)	0.959 (0.944, 0.973)	−0.014 (−0.029, 0.001)	−0.041 (−0.063, −0.022)
(n = 941,	LR+ (95% CI)	4.77 (4.08, 5.61)	4.42 (3.78, 5.19)	0.35 (−0.23, 0.84)	14.1 (10.6, 19.0)	−9.4 (−13.9, −6.3)	−9.7 (−14.6, −6.7)
7 centers)	LR− (95% CI)	0.09 (0.06, 0.14)	0.14 (0.10, 0.19)	−0.05 (−0.08, −0.01)	0.07 (0.05, 0.11)	0.02 (−0.01, 0.05)	0.07 (0.02, 0.11)
	Sensitivity (%)	92.7	89.2	3.5 (0.7, 6.6)	93.0	−0.3 (−3.5, 2.7)	−3.8 (−7.8, −0.1)
	Invasive tumors only (%)	96.8	95.4	1.4 (−1.4, 4.5)	96.3	0.5 (−2.2, 3.2)	−0.9 (−4.4, 2.4)
	Specificity (%)	80.6	79.8	0.8 (−1.2, 2.8)	93.4	−12.8 (−15.8, −10.1)	−13.6 (−16.7, −10.7)
	Predicted vs. observed risk*	0.84	0.81				

95% CIs for area under the receiver–operating characteristics curve (AUC), differences in AUC and differences in LR+ and LR− were based on the bias-corrected bootstrap method using 1000 bootstrap samples; 95% CIs for LR+ and LR− were based on the Cox–Hinkley–Miettinen–Nurminen method[18]; 95% CIs for sensitivity and specificity differences were obtained using a method based on continuity-corrected Wilson score intervals (method ten from Justice *et al.*[17]). *Ratio between average predicted probability of malignancy and observed prevalence of malignancy.

cancers and three were rare malignancies)) and gave 107 false-positive results, resulting in an LR+ of 6.36 and an LR− of 0.10. Thus, LR1 provided a classification advantage over LR2 of one in 255 malignancies (0.4%) and seven out of 742 benign tumors (0.9%). Subjective assessment missed 32 cancers (nine borderline tumors, two metastatic tumors in the ovary and 21 primary invasive malignancies (of which three were Stage I epithelial cancers, three were Stage II epithelial cancers, eight were Stage III epithelial cancers and seven were rare malignancies)) and produced 59 false-positive results. This corresponds to an LR+ of 11.0 and an LR− of 0.14. Thus, LR1 provided a classification advantage over subjective assessment of 11 out of 255 malignancies (4.3%) and a classification disadvantage in 46 of the 742 benign tumors (6.2%).

Overall, the risk of malignancy was slightly underpredicted by LR1 (ratio of predicted and observed risk, 0.83) and LR2 (ratio of predicted and observed risk, 0.78). This underestimation can be seen in the calibration curves (Figure S1).

Temporal validation

The diagnostic performance of LR1, LR2 and subjective assessment in the seven old centers (n = 941) is shown in Tables 4, S1 and S2 and in Figures 2 and S1. The AUCs of the three methods were 0.945, 0.918 and 0.959, respectively. In the old centers the AUC for LR2 was

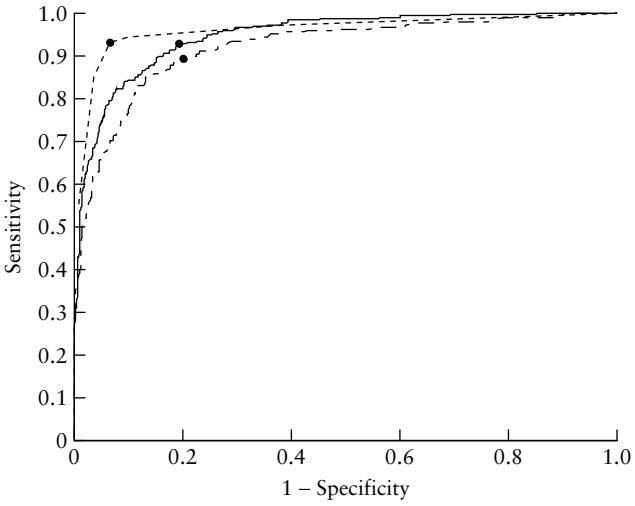

Figure 2 Receiver–operating characteristics (ROC) curves for logistic regression models LR1 (——) and LR2 (– - –), and subjective assessment using pattern recognition (- - - - -) on temporal validation (n = 941). The dots in the ROC curves for LR1 and LR2 correspond to the sensitivity and false-positive rate (1 minus specificity) for the thresholds (i.e. an estimated risk of malignancy of 0.10) recommended in the study in which the models were created[12]. The ROC curve for subjective assessment was created using six levels of diagnostic confidence. The dot in the ROC curve for subjective assessment corresponds to sensitivity and false-positive rate of the ultrasound examiner's dichotomous classification of the tumor as benign or malignant. Areas under the curves were 0.945 for LR1, 0.918 for LR2 and 0.959 for pattern recognition.

smaller than in the new centers, the specificities for LR1 and LR2 were lower and the sensitivity for subjective assessment was higher.

DISCUSSION

On external validation we showed that the IOTA logistic regression models can predict the presence of ovarian malignancy in women with an adnexal mass and that the performance of the models is equivalent to subjective assessment by gynecologists and radiologists specialized in gynecological ultrasound examinations and who have a special interest in adnexal tumors. This is the first prospective study to externally validate the performance of IOTA logistic regression models to distinguish between benign and malignant adnexal masses. The simpler model containing six variables (LR2) performed almost as well as the model containing 12 variables (LR1). This is encouraging because a model with a small number of variables is likely to be more user-friendly.

It is a strength that the models were both developed and tested in a large multicenter study and that data were obtained using a variety of ultrasound machines. This is likely to make our results generalizable, and the results of our external validation suggest that the models are robust. However, the fact that the models were both developed and tested only by experienced ultrasound examiners could make them less generally applicable. However, we believe that it should be possible for any qualified ultrasound practitioner to obtain information on the ultrasound variables required for these models. A potential weakness of our study is that only masses that underwent surgery were included. This is inevitable, as ideally histology should be used as a gold standard in a study estimating sensitivity and specificity with regard to malignancy. Moreover, masses not requiring surgery are likely to be less complex and easier to classify.

Both models performed very well in old and new centers. Surprisingly, the LR+ was higher on external validation in comparison with temporal validation, whereas the LR− was the same in both groups. In the new centers (external validation) we found more teratomas, simple cysts and functional cysts, which are 'easier' tumors to classify, and fewer Stage I tumors and borderline malignant tumors, which are more difficult to classify[26] (Table 2). This may have explained the better LR+ and specificity on external validation. Both the LR1 and LR2, and subjective assessment, had previously undergone temporal validation in three of the old centers[13]. The results obtained were very similar to those presented in this work, of the temporal validation carried out in seven old centers, supporting that the models are robust.

The serum marker CA 125 has been used as an indicator of ovarian cancer[27,28]. For example, the risk of malignancy index[10] can identify the benign or malignant nature of a mass (AUC = 0.87, LR+ = 3.49 and LR− = 0.27), but it does so less well than LR1 (AUC = 0.94, LR+ = 3.61 and LR− = 0.10) and LR2 (AUC = 0.92, LR+ = 3.07 and LR− = 0.14)[15]. We have demonstrated that subjective assessment is superior to single measurements of serum CA 125 for discriminating between benign and malignant adnexal masses[29], and that CA 125 does not add any diagnostic information when an ovarian mass is examined using ultrasound techniques by an experienced examiner[14]. We have also shown that the inclusion of serum CA 125 results in logistic regression models does not improve the prediction of malignancy in adnexal masses[30].

Subjective assessment has been shown to be the best method for classifying adnexal masses as benign or malignant[12–14], even though this requires a level of experience that not all ultrasound examiners have. As shown in this study, our models perform very well in the hands of experienced ultrasound examiners. However, our models are likely to be more helpful to clinicians who are not ultrasound experts, and to sonographers, who generally describe the features of a mass rather than suggesting a diagnosis. It remains to be shown whether the models work also in their hands. The models could be used in any environment because they are freely available and can be installed on any personal computer or even in an ultrasound system. On external validation the sensitivity for primary invasive cancer of the models was higher than that of subjective assessment. Both models missed 6% of invasive cancers and subjective assessment missed 11%. From a patient's perspective this is important. Maximizing sensitivity is more important than maximizing specificity. The problem with subjective assessment is that there is no clear cut-off so the sensitivity cannot be easily improved. This would suggest that using the models instead of subjective assessment may lead to fewer ovarian cancers being operated on inappropriately by general gynecologists. This is advantageous, because the long-term prognosis of patients with ovarian cancer is better if an oncological surgeon performs the primary operation, and patients with early stage ovarian cancer who are treated in specialized and semi-specialized hospitals live for longer than patients treated in general hospitals[31]. If the correct terms, definitions and measurements[21] are used by qualified examiners, the models could make a significant improvement to the management of women with suspected ovarian pathology.

ACKNOWLEDGMENTS

This research was supported by Research Council KULeuven: GOA-MANET, CoE EF/05/006 Optimization in Engineering (OPTEC); FWO: G.0302.07 (SVM), research communities (ICCoS, ANMMM); IWT-TBM 070706 (IOTA); Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO); EU: BIOPATTERN (FP6-2002-IST 508803); 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). Ben Van Calster is a postdoctoral researcher funded by the Research Foundation – Flanders (FWO), Belgium. We thank all

participating centers, the principal investigators and the study participants for their contribution.

APPENDIX

Recruitment centers

University Hospitals Leuven (Belgium); Ospedale S. Gerardo, Università di Milano Bicocca, Monza (Italy); Ziekenhuis Oost-Limburg (ZOL), Genk, (Belgium); Medical University in Lublin (Poland); University of Cagliari, Ospedale San Giovanni di Dio, Cagliari (Italy); Malmö University Hospital, Lund University (Sweden); University of Bologna (Italy); Università Cattolica del Sacro Cuore Rome (Italy); DCS Sacco University of Milan (Milan A) (Italy); General Faculty Hospital of Charles University, Prague (Czech Republic); Chinese PLA General Hospital, Beijing (China); King's College Hospital London (UK); Universita degli Studi di Napoli, Napoli (Naples A) (Italy); IEO, Milano (Milan B) (Italy); Lund University Hospital, Lund (Sweden); Macedonio Melloni Hospital, University of Milan (Milan C) (Italy); Università degli Studi di Udine (Italy); McMaster University, St Joseph's Hospital, Hamilton, Ontario (Canada); and Instituto Nationale dei Tumori, Fondazione Pascale, Napoli (Naples B) (Italy)

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

IOTA principal investigators (alphabetical order)

Artur Czekierdowski, Lublin, Poland
Elisabeth Epstein, Lund, Sweden
Daniela Fischerová, Prague, Czech Republic
Dorella Franchi, Milano, Italy
Robert Fruscio, Monza, Italy
Stefano Greggi, Napoli, Italy
Stefano Guerriero, Cagliari, Italy
Jingzhang, Beijing, China
Davor Jurkovic, London, UK
Francesco P.G. Leone, Milano, Italy
Andrea A. Lissoni, Monza, Italy
Henry Muggah, Hamilton, Ontario, Canada
Dario Paladini, Napoli, Italy
Alberto Rossi, Udine, Italy
Luca Savelli, Bologna, Italy
Antonia Carla Testa, Roma, Italy
Dirk Timmerman, Leuven, Belgium
Diego Trio, Milano, Italy
Lil Valentin, Malmö, Sweden
Caroline Van Holsbeke, Genk, Belgium

REFERENCES

1. Moore RG, McMeekin DS, Brown AK, DiSilvestro P, Miller MC, Allard WJ, Gajewski W, Kurman R, Bast RC Jr, Skates SJ. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol Oncol* 2009; **112**: 40–46.
2. Cannistra S. Ovarian cancer. *N Engl J Med* 2004; **351**: 2519–2529.
3. 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.
4. Medeiros LR, Stein AT, Fachel J, Garry R, Furness S. Laparoscopy versus laparotomy for benign ovarian tumours. *Cochrane Database Syst Rev* 2005; **20**: CD004751.
5. 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.
6. 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.
7. Bourne T, Campbell S, Steer C, Whitehead MI, Collins WP. Transvaginal colour flow imaging: a possible new screening technique for ovarian cancer. *BMJ* 1989; **299**: 1367–1370.
8. 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.
9. 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.
10. 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.
11. 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. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the International Ovarian Tumor Analysis (IOTA) Group. *Clin Cancer Res* 2007; **13**: 4440–4447.
12. 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.
13. Van Holsbeke C, Van Calster B, Testa AC, Domali E, Lu C, Van Huffel S, Valentin L, Timmerman D. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the International Ovarian Tumor Analysis Study. *Clin Cancer Res* 2009; **15**: 684–691.
14. Valentin L, Jurkovic D, Van Calster B, Testa A, Van Holsbeke C, Bourne T, Vergote I, Van Huffel S, Timmerman D. Adding a single CA-125 measurement to ultrasound performed by an experienced examiner does not improve preoperative discrimination between benign and malignant adnexal masses. A prospective international multicentre study of 809 patients. *Ultrasound Obstet Gynecol* 2009; **34**: 345–354.
15. 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.

16. Wyatt JC, Altman DG. Prognostic models: clinically useful or quickly forgotten? *BMJ* 1995; **311**: 1539–1541.
17. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; **130**: 515–524.
18. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statist Med* 2000; **19**: 453–473.
19. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009; **338**: b605.
20. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HC. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003; **138**: 40–44.
21. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. 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.
22. Heintz AP, Odicino F, Maisonneuve P, Quinn MA, Benedet JL, Creasman WT, Ngan HY, Pecorelli S, Beller U. Carcinoma of the ovary. FIGO 6th Annual Report on the Results of Treatment in Gynecological Cancer. *Int J Gynaecol Obstet* 2006; **95** (Suppl. 1)**:** S161–S192.
23. Steyerberg EW. *Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating*. Springer: New York, 2009.
24. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *BMJ* 2009; **339**: b2584.
25. Cleveland WS, Devlin SJ, Grosse E. Regression by local fitting. *J Econometrics* 1988; **37**: 87–114.
26. 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.
27. Bast RC Jr, Klug TL, St John E, Jenison E, Niloff JM, Lazarus H, Berkowitz RS, Leavitt T, Griffiths CT, Parker L, Zurawski VR Jr, Knapp RC. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *N Engl J Med* 1983; **309**: 883–887.
28. Einhorn N, Bast RC Jr, Knapp RC, Tjernberg B, Zurawski VR Jr. Preoperative evaluation of serum CA125 levels in patients with primary epithelial ovarian cancer. *Obstet Gynecol* 1986; **67**: 414–416.
29. Van Calster B, Timmerman D, Bourne T, Testa AC, Van Holsbeke C, Domali E, Jurkovic D, Neven P, Van Huffel S, Valentin L. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125. *J Natl Cancer Inst* 2007; **99**: 1706–1714.
30. Timmerman D, Van Calster B, Jurkovic D, Valentin L, Testa AC, Bernard JP, Van Holsbeke C, Van Huffel S, Vergote I, Bourne T. Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors. *J Clin Oncol* 2007; **25**: 4194–4200.
31. Vernooij F, Heintz APM, Witteveen PO, van der Heiden-van der Loo M, Coebergh JW, van der Graaf Y. Specialized care and survival of ovarian cancer patients in the Netherlands: nationwide cohort study. *J Natl Cancer Inst* 2008; **100**: 399–406.

SUPPORTING INFORMATION ON THE INTERNET

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

Figure S1 Calibration curves for LR1 and LR2 on both temporal ($n = 941$) and external ($n = 997$) validation. The diagonal line represents the ideal situation; the predicted risk and observed prevalence of malignancy are identical.

Table S1 Area under the receiver–operating characteristics curves (AUC) of LR1 and LR2 in pre- and postmenopausal women.

Table S2 Area under the receiver–operating characteristics curves (AUC) of logistic regression models LR1 and LR2 in each center.

