VOLUME 23 · NUMBER 34 · DECEMBER 1 2005

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group

Dirk Timmerman, Antonia C. Testa, Tom Bourne, Enrico Ferrazzi, Lieveke Ameye, Maja L. Konstantinovic, Ben Van Calster, William P. Collins, Ignace Vergote, Sabine Van Huffel, and Lil Valentin

From the Department of Obstetrics and Gynecology, University Hospitals Katholieke Universiteit Leuven; Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium; Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Rome; Dipartimento di Scienze Cliniche, Sacco, Università di Milano, Milan, Italy; Department of Obstetrics and Gynaecology, St George's Hospital Medical School, University of London; King's College London, London, United Kingdom; and Department of Obstetrics and Gynecology, University Hospital Malmö, Malmö Sweden.

Submitted February 25, 2005; accepted September 7, 2005.

Supported by interdisciplinary research grants of the research council of the Katholieke Universiteit Leuven, Belgium (IDO/99/03 and IDO/02/09 projects), by the Belgian Programme on Interuniversity Poles of Attraction (phase V-22), by the Concerted Action Project AMBioRICS of the Flemish Community, by the FWO project G.0407.02, by the European Union Network of Excellence BIOPATTERN (contract No. FP6-2002-IST 508803), by research grants from the Swedish Medical Research Council (Grants No. K98-17X - 11,605 - 03A, K2001 - 72X - 11,605 - 06A and K2002- 72X- 11,605- 07B), by funds administered by Malmö University Hospital, Allmänna Sjukhusets i Malmö Stiftelse för bekämpande av cancer (the Malmö General Hospital Foundation for fighting against cancer), and ALF-medel (a Swedish governmental grant from the region of Scania).

Presented in part at an invited lecture at the 14th World Congress of Ultrasound in Obstetrics and Gynecology, organized by the International Society of Ultrasound in Obstetrics and Gynecology, August 31-September 4, 2004, Stockholm, Sweden.

Authors' disclosures of potential conflicts of interest are found at the end of this article.

Address reprint requests to Dirk Timmerman, MD, PhD, Department of Obstetrics and Gynecology, University Hospitals, Katholieke Universiteit Leuven, Herestraat 49, B-3000 Leuven, Belgium; e-mail: dirk.timmerman@uz.kuleuven.ac.be.

© 2005 by American Society of Clinical Oncology

0732-183X/05/2334-8794/$20.00

DOI: 10.1200/JCO.2005.01.7632

ABSTRACT

Purpose
To collect data for the development of a more universally useful logistic regression model to distinguish between a malignant and benign adnexal tumor before surgery.

Patients and Methods
Patients had at least one persistent mass. More than 50 clinical and sonographic end points were defined and recorded for analysis. The outcome measure was the histologic classification of excised tissues as malignant or benign.

Results
Data from 1,066 patients recruited from nine European centers were included in the analysis; 800 patients (75%) had benign tumors and 266 (25%) had malignant tumors. The most useful independent prognostic variables for the logistic regression model were as follows: (1) personal history of ovarian cancer, (2) hormonal therapy, (3) age, (4) maximum diameter of lesion, (5) pain, (6) ascites, (7) blood flow within a solid papillary projection, (8) presence of an entirely solid tumor, (9) maximal diameter of solid component, (10) irregular internal cyst walls, (11) acoustic shadows, and (12) a color score of intratumoral blood flow. The model containing all 12 variables (M1) gave an area under the receiver operating characteristic curve of 0.95 for the development data set ($n = 754$ patients). The corresponding value for the test data set ($n = 312$ patients) was 0.94; and a probability cutoff value of .10 gave a sensitivity of 93% and a specificity of 76%.

Conclusion
Because the model was constructed from multicenter data, it is more likely to be generally applicable. The effectiveness of the model will be tested prospectively at different centers.

J Clin Oncol 23:8794-8801. © 2005 by American Society of Clinical Oncology

INTRODUCTION

The preoperative assessment of adnexal tumors remains a major challenge for the gynecologist. Advances in surgery have provided more treatment options, but their potential usefulness depends on a prior assessment of the mass using noninvasive procedures. Appropriate surgical treatment is essential because the rupture of a stage 1 ovarian cancer during the operation may worsen the prognosis.[1]

In 1989, Granberg et al[2] reported that transvaginal sonographic images of an ovarian mass could be used to predict the likelihood of malignancy. This approach was later incorporated into scoring systems to improve test performance.[3-5] In 1989, it was also shown that an ultrasound-derived index of tumoral blood flow might be an important risk factor.[6] Other workers used menopausal status and the serum CA-125 level to produce a risk-of-malignancy index.[7] The logical extension of this work was to use logistic regression models[8,9] or neural networks[10,11] to establish the most important variables and improve test performance. However, the results of prospective evaluations of these models at different centers were not encouraging.[12-14] One study reported values for sensitivity and specificity figures as low as 62% and 79%, respectively,[14] whereas another study reported unbalanced figures of 9% and 91%, respectively.[12] A retrospective assessment of all reports suggests that the study populations were insufficiently large for model development or evaluation because of differences in the prevalence of different types of ovarian malignancies and rare benign tumors. There were also variations in the interpretation of ultrasound-derived end points.

The aim of the International Ovarian Tumor Analysis Group study was to minimize the limitations of previous reports by studying a total of 1,000 patients with a persistent adnexal mass at a minimum of six centers working with a common protocol. More than 50 defined variables were recorded on a database for the development of a logistic regression model that could be used to achieve high sensitivity ($> 90\%$) for malignant ovarian tumors (and a specificity $> 75\%$).

The multicenter approach was chosen as the most likely method to achieve a model that might be more effectively applicable to prospective studies in different clinic populations.

PATIENTS AND METHODS

The study was prospective and multicenter. The protocol was ratified by the local ethics committee at each recruitment center.

Recruitment Centers

There were nine centers from five countries: University Hospital Malmö (Sweden), University Hospital Leuven (Belgium), Universita del Sacro Cuore Rome (Italy), Dipartimento di Scienze Cliniche, Sacco University of Milan (Italy), Hôpital Boucicaut Paris (France), Centre Medical des Pyramides, Maurepas (France), King's College Hospital London (United Kingdom), Istituto di Scienze Biomediche Ospedale, San Gerado Universita di Milano, Monza (Italy), and Universita degli Studi di Napoli, Naples (Italy).

Patients

Inclusion criteria were as follows: patients presenting with at least one overt persistent adnexal mass who were assessed by a principal investigator at one of the participating centers were eligible for inclusion in the study. Data from the apparent worse case mass were used for this study.

Patients were excluded from study for the following reasons: pregnancy or refusal of transvaginal sonography, surgery more than 120 days after sonographic assessment, disagreement in the classification (malignant or benign) between the original pathology report and the report of an expert reviewer, or incomplete submission of the data.

Data Collection

A dedicated, secure data collection system was developed for the study.[15] A unique identifier was generated automatically for each patient's record based on the identifier for the center, the patient's birthday, the date of the ultrasound scan, and the follow-up number. Clinicians at each center could only view or update patient records from their own center. Data security was ensured by not recording the patient's name and by encrypting all data communication between the browser-based data entry and the central database using a 56-bit SSL (Secure Socket Layer) certificate. Data integrity was ensured by client-side JavaScript checks, server-side XML-encoded rules based on the study protocol, and manual checks by three experts.

Clinical Variables

A family history that included the number of first-degree relatives with ovarian or breast cancer was taken from each patient. Demographic data included the patient's age, menopausal status, day of menstrual cycle (if appropriate), previous hormonal therapy, and surgical history. Women ≥ 50 years of age who had undergone a hysterectomy were defined as postmenopausal.

Sonographic End Points

A transvaginal scan was performed in all cases. Transabdominal sonography was used to examine a large mass that could not be seen in its entirety using a transvaginal probe. A standardized sonographic procedure was used at all centers. Gray scale and color Doppler images were used to obtain over 40 morphologic and blood flow end points to characterize each adnexal mass. These criteria have been illustrated, described and defined.[16] When intratumoral blood flow velocity waveforms were not detected, the peak systolic velocity, time averaged maximum velocity, the pulsatility index, and the resistance index were coded as 2.0 cm/sec, 1 cm/sec, 3.0 cm/sec, and 1.0 cm/sec, respectively, for use in mathematical modeling. The presence or absence of pain during the examination was recorded. Finally, the investigator gave a subjective assessment of whether the mass was likely to be malignant or benign.

Serum Tumor Marker

Centers were 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 immunoradiometric assay CA-125 II (Centocor, Malvern, PA; or Cis-Bio, Gif-sur-Yvette, France; or Abbott Axsym system, REF 3B41-22, Abbott Laboratories Diagnostic Division, Abbott Park, IL; or Immuno-l-analyser; Bayer Diagnostics, Tarrytown, NY; or Vidas; bioMétrieux, Marcy l'Etoile, France) was used and the results expressed in units per milliliter.

Outcome Measures

The final outcome measures of the study were the histologic diagnosis and, in the cases of malignancy, the surgical stage. Surgery was performed in the case of a mass classified as persistent (ie, still present 6 to 12 weeks after the initial scan). In cases of symptomatic masses, suspected malignancy, or at the patient's request, surgery was performed more quickly, either by laparoscopy or laparotomy according to the surgeon's judgment. All excised tissues were sampled for histologic examination at the local center. Tumors were classified according to the criteria recommended by the International Federation of Gynecology and Obstetrics.[17] The degree of differentiation of malignant tumors was recorded. The pathologic samples from approximately 10% of the patients were randomly selected for peer review by Professor Ph. Moerman (Katholieke Universiteit, Leuven, Belgium).

Statistical Analysis

The data were thoroughly inspected using univariate analyses (contingency tables and basic descriptive statistics) and multivariate analyses (principal component analysis, bi-plots, scatter matrices, canonical correlation analysis, and logistic regression) to explore the data and to detect multicollinearity and outliers. Manual checks of any outliers were performed to eliminate mistakes that had occurred during submission of the data. To address the problem of characterizing preoperatively a mass as malignant or benign, we randomly stratified 70% of patient data to construct the logistic regression model and 30% for use as the test data. The data were stratified 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. Then based on forward-backward selection methods (using the statistical significance level of the differences in χ^2) and after checking for interactions among the factors and determining whether the model was linear in the logit for continuous factors, an input selection was made for logistic regression models.[18-20] Receiver operating characteristic (ROC) curves were constructed, and the method proposed by DeLong et al[21] was used to check for statistically significant differences in the area under the ROC curves.

RESULTS

A total of 1,149 patients were recruited. Data from 83 patients (7%) were excluded (eight because of pregnancy, 31 because surgery was undertaken more than 120 days from the sonographic assessment, 42 because of incomplete submission of data, and two because of disagreement over the histologic diagnosis). Most missing data were the result of missing pathology owing to no operation being performed. Data from 1,066 patients (93%) were available for statistical analysis and model development.

The overall mean age of the patients was 47 years (range, 17 to 94 years), the overall proportion of patients who were postmenopausal was 40.5%, the overall proportion who were nulliparous was 37.7%, and the overall proportion receiving hormonal therapy was 22.1% (Table 1). There were a total of 52 stage I primary ovarian cancers.

Univariate Analysis

The outcome of a univariate analysis of continuous demographic data and ultrasound-based variables is shown in Table 2. The level of serum CA-125 was measured in 70.9% of patients with a benign tumor and in 91.0% of patients with a malignant lesion. All variables with the exception of years since menopause were significantly different between malignant and benign tumors. An analysis of all binary end points is shown in Table 3. An analysis of all categoric end points is shown in Table 4. All of these variables were significantly different between malignant and benign tumors.

Only two patients presented with a single unilocular tumor that proved to be malignant: one simple cyst (140 × 115 × 105 mm) with color score 2 in a 26-year-old patient with a serum CA-125 level of 13 U/mL proved to be a borderline malignant serous papillary cystadenoma, stage Ia; and one simple cyst (55 × 38 × 35 mm) with color score 1 in a 42-year-old patient with a serum CA-125 level of 7 U/mL proved to be a borderline malignant mucinous cystadenoma, stage Ia.

Logistic Regression Analysis

Stepwise multivariate regression analysis of the development data set (754 patients) resulted in the delineation of the optimal set of variables that could be included in the equation: (1) personal history of ovarian cancer (yes = 1, no = 0), (2) current hormonal therapy (yes = 1, no = 0), (3) age of the patient (in years), (4), maximum diameter of the lesion (in millimeters), (5) the presence of pain during the examination (yes = 1, no = 0), (6) the presence of ascites (yes = 1, no = 0), (7) the presence of blood flow

Table 1. Histologic Outcome of Adnexal Tumors by Participating Center

	Total		Benign		Malignant		Primary Invasive		Borderline		Metastatic	
Center	No.	%*	No.	%†	No.	%†	No.	%‡	No.	%‡	No.	%‡
Malmö	315	29.5	247	78.4	68	21.6	40	58.8	17	25.0	11	16.2
Leuven	263	24.7	170	64.6	93	35.4	62	66.7	14	15.1	17	18.3
Rome	126	11.8	81	64.3	45	35.7	23	51.1	12	26.7	10	22.2
Milano	87	8.2	79	90.8	8	9.2	6	75.0	1	12.5	1	12.5
Paris§	80	7.5	71	88.8	9	11.2	7	77.8	2	22.2	0	0
Paris‖	64	6.0	57	89.1	7	10.9	6	85.7	1	14.3	0	0
London	54	5.1	38	70.4	16	29.6	10	62.5	4	25.0	2	12.5
Monza	46	4.3	29	63.0	17	37.0	12	70.6	4	23.5	1	5.9
Naples	31	2.9	28	90.3	3	9.7	3	100.0	0	0	0	0
All	1,066	100	800	75.0	266	25.0	169	63.5	55	20.7	42	15.8

*Of study total.
†Of center total.
‡Of malignant tumors.
§Paris, Baucicout.
‖Paris, Maurepas.

Table 2. Univariate Analysis of Most Continuous End Points After the Histologic Classification of Adnexal Tumors As Benign or Malignant

	Benign				Malignant			
Variable	No.	Median	Minimum	Maximum	No.	Median	Minimum	Maximum
Demographic								
Age, years*	800	42	17	90	266	56	17	94
Years postmenopause†	229	10	1	40	145	12	0	44
Parity*	800	1	0	10	266	2	0	7
Morphologic								
Max Les Dia, mm*	800	63	11	320	266	100.5	8	410
Volume Les, mL*	800	73	0.2	7,781	266	303	0.1	11,829
Fluid in P of D, mm*	140	12	2	61	144	24	3	100
Septum, mm*	343	2.1	1	20	143	4.0	1	20
Pap ht, mm*	156	7	2	62	121	14	3	62
Max Pap, mm*	156	10	3	90	121	21	4	110
Ratio Pap Les††	156	0.003	0	0.456	121	0.006	0	0.420
Pap Nr*	156	1	1	> 3	121	> 3	1	> 3
Loc Nr*	800	1	0	> 10	266	3	0	> 10
Max Solid Dia, mm*	309	21	3	230	244	50	4	214
Solid Vol, mL*	309	1.6	0.006	1,978	244	34	0.008	2,291
Ratio Solid Les*	309	0.03	0	1	244	0.24	0	1
Bloodflow								
PI*	506	0.95	0.13	5.80	246	0.74	0.25	2.26
RI*	506	0.59	0.12	1.0	246	0.50	0.17	1.0
PSV*	506	11.4	2.0	85.5	246	24.30	3.9	202
TAMXV*	498	6.9	1.0	60.0	241	17.0	3.0	137
Tumor marker CA-125 (U/mL)	567	17	1.0	1,409	242	167	4	31,610

Abbreviations: Max Les D, maximal diameter of the lesion; Volume Les, volume of the lesion; Fluid in P of D, fluid in anterioposterior plane of pouch of Douglas; Septum, thickness; Pap ht, height of papillary structure; Max Pap, maximal diameter of papillary structure; Ratio Pap Les, ratio between volume of the papillary structure and volume of the lesion; Pap Nr, number of separate papillary projections (1, 2, 3, or > 3); Loc Nr, number of locules (0, 1, 2, 3, 4, 5 to 10, or > 10); Max Solid Dia, maximal diameter of the solid component; Solid Vol, volume of the largest solid component; Ratio Solid Les, ratio between volume of the largest solid component and volume of the lesion; PI, pulsatility index; RI, resistance index; PSV, peak systolic velocity (cm/s); TAMXV, time-averaged maximum velocity; CA-125, serum level of the tumor marker.

*$P < .0001$.

†$P = .14$.

‡$P = .0055$ (Mann-Whitney).

within a solid papillary projection (yes = 1, no = 0), (8) the presence of a purely solid tumor (yes = 1, no = 0), (9) maximal diameter of the solid component (expressed in millimeters, but with no increase > 50 mm), (10) irregular internal cyst walls (yes = 1, no = 0), (11) the presence of acoustic shadows (yes = 1, no = 0), and (12) the color score (1, 2, 3, or 4).

The logistic regression model (M1) provided the estimated probability of malignancy for a particular patient with an adnexal tumor. This probability was equal to $y = 1/(1 + e^{-z})$, where $z = -6.7468 + 1.5985\,(1) - 0.9983\,(2) + 0.0326\,(3) + 0.00841\,(4) - 0.8577\,(5) + 1.5513\,(6) + 1.1737\,(7) + 0.9281\,(8) + 0.0496\,(9) + 1.1421\,(10) - 2.3550\,(11) + 0.4916\,(12)$, and e is the mathematical constant and base value of natural logarithms.

Independent risk factors for the presence of malignancy included the patient's age (a 3.3% increase of odds for malignancy with each additional year of age), a personal history of ovarian cancer (odds ratio, 4.95), the maximum diameter of the lesion (0.8% for every millimeter increment), the maximum diameter of the solid component (5.1% for every millimeter increase with an upper limit of 50 mm; no increase > 50 mm), the presence of ascites (odds ratio, 4.72), the presence of blood flow within a solid papillary projection (odds ratio, 3.23), the presence of a purely solid lesion (odds ratio, 2.53), the presence of irregular internal cyst walls (odds ratio, 3.13), increase in color score (odds ratio, 1.64 for every one unit increase). Factors that reduced the risk of malignancy included the current use of hormonal therapy (odds ratio, 0.369), the presence of pain during the ultrasound examination (odds ratio, 0.424), and the presence of acoustic shadows (odds ratio, 0.095).

A simpler version (M2) using six selected variables was also developed. These were the six variables that were first entered into the model M1 when using stepwise selection of variables. The variables included in the equation M2 were (1) age of the patient (in years), (2) the presence of ascites (yes = 1, no = 0), (3) the presence of blood flow within a solid papillary projection (yes = 1, no = 0), (4) maximal diameter of the solid component (expressed in millimeters, but with no increase > 50 mm), (5) irregular internal cyst walls (yes = 1, no = 0), and (6) the presence of acoustic

Table 3. Univariate Analysis of All Binary End Points After the Histologic Classification of Adnexal Tumors As Benign or Malignant

	Benign		Malignant		
	No.	%	No.	%	*P*
Fam His Ov Ca	800	2.5	266	4.9	.0559
Fam His Br Ca	800	10.8	266	12.4	.4578
Pers His Ov Ca	800	0.8	266	3.0	.0096
Pers His Br Ca	800	2.9	266	5.6	.0386
Nullipara	800	41.9	266	25.2	< .0001
Hysterectomy	800	6.3	266	10.5	.0216
Postmenopause	800	32.9	266	63.5	< .0001
Hormonal therapy	800	23.5	266	17.7	.0477
PM bleeding	229	14.9	145	17.9	.4291
Bilateral masses	800	16.6	266	31.2	< .0001
Pelvic pain	800	28.8	266	19.6	.0034
Ascites	800	2.9	266	42.1	< .0001
Incomp septum	800	9.3	266	4.1	.0094
Papillation	800	19.5	266	45.5	< .0001
Pap blood flow	156	33.3	121	84.3	< .0001
Pap smooth, irregular	156	49.4	121	82.6	< .0001
Internal wall, irregular	800	32.8	266	81.6	< .0001
Acoustic shadows	800	13.0	266	1.5	< .0001
Venous blood flow	800	9.0	266	3.4	.0040

Abbreviations: Fam His Ov Ca, number of women with first-degree relatives with ovarian cancer; Fam His Br Ca, number of women with first-degree relatives with breast cancer; Pers His Ov Ca, personal history of ovarian cancer; Pers His Br Ca, personal history of breast cancer; Nullipara, never gave birth; PM bleeding, presence of postmenopausal bleeding within the year before ultrasound examination; Bilateral masses, presence of a lesion on both adnexal sides; Pelvic pain, presence of pain during the examination; Ascites, fluid outside the pouch of Douglas; Incomp septum, presence of an incomplete septum; Pap blood flow, presence of flow within at least one of the papillary projections; Pap smooth (irregular), presence of an irregular papillary projection; Internal wall, irregular, presence of irregular internal walls in the lesion; acoustic shadows, presence of acoustic shadows; venous blood flow, no arterial blood flow detected, but venous flow only.

Table 4. Univariate Analysis of All Categorical End Points After the Histologic Classification of Adnexal Tumors As Benign (n = 800) or Malignant (n = 266)

	Benign		Malignant	
Variable	No.	%	No.	%
Locularity*				
Unilocular	311	38.9	2	0.8
Unilocular, solid	88	11.0	44	16.5
Multilocular	176	22.0	20	7.5
Multilocular, solid	168	21.0	116	43.6
Solid	52	6.5	84	31.6
Not classifiable	5	0.6	0	0
Echogenicity*				
Anechogenic	303	37.9	107	40.2
Low level	149	18.6	60	22.6
Ground glass appearance	192	24.0	33	12.4
Hemorrhagic	8	1.0	2	0.8
Mixed echogenicity	114	14.3	18	6.8
No cyst fluid	34	4.3	46	17.3
Color Score, blood flow*				
No flow (1)	222	27.8	11	4.1
Minimal flow (2)	311	38.9	42	15.8
Moderately strong flow (3)	219	27.4	109	41.0
Very strong flow (4)	48	6.0	104	39.1

*$P < .0001$.

shadows (yes = 1, no = 0). For the reduced model the probability was equal to $y = 1/(1 + e^{-z})$, where $z = -5.3718 + 0.0354\ (1) + 1.6159\ (2) + 1.1768\ (3) + 0.0697\ (4) + 0.9586\ (5) - 2.9486\ (6)$.

Model Evaluation

The final model (M1) applied to the development data set gave the ROC shown in Figure 1. The area under the curve (AUC) was 0.95 (SE = 0.009), and the corresponding value for model (M2) was 0.93 (SE = 0.012). The numbers of malignant and benign tumors that were correctly or incorrectly classified using M1 is shown against the use of different probability levels in Table 5. A probability value of .10 gave a sensitivity of 93% and a specificity of 77%. The corresponding values for M2 at a probability level of .10 were 92% and 75%, respectively.

The results of applying model M1 to the test data set (312 patients) are summarized in Table 6. The ROC from applying the model M1 to the test data set is also shown in Figure 1. The AUC was 0.94 (SE = 0.017), and a probability value of .10 gave a sensitivity of 93% and a specificity of 76%. The corresponding values for M2 were 0.92 (SE = 0.018) for AUC, 89% for sensitivity, and 73% for specificity. Model M1 performed significantly better than M2 on the test data set when AUCs were compared ($P = .028$). When the model was tested on the data from each individual center, the overall test performance did not deteriorate. In Table 7, the results of applying different older models to the test data set are compared with the

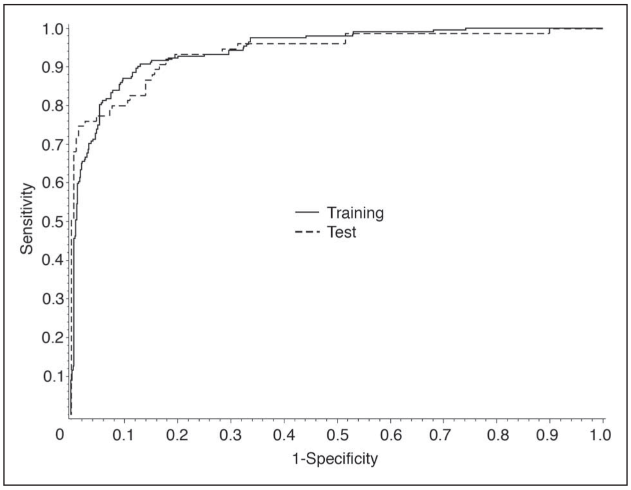

Fig 1. The receiver operating characteristic curves of the logistic regression model constructed on the development data set (n = 754) and applied to the test set (n = 312). The areas under the curve are 0.946 and 0.942, respectively.

Table 5. Classification of Malignant and Benign Tumors by Probability Level After Development of a Logistic Regression Model From the Development Data Set (n = 754)

	Correctly Classified		Incorrectly Classified						
Probability Level (*P*)	Malignant	Benign	Malignant As Benign	Benign As Malignant	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
.01	189	149	2	414	99.0	26.5	44.8	31.3	98.7
.05	180	364	11	199	94.2	64.7	72.1	47.5	97.1
.10	177	433	14	130	92.7	76.9	80.9	57.7	96.9
.15	173	464	18	99	90.6	82.4	84.5	63.6	96.3
.20	171	481	20	82	89.5	85.4	86.5	67.6	96.0
.25	166	496	25	67	86.9	88.1	87.8	71.2	95.2
.30	165	504	26	59	86.4	89.5	88.7	73.7	95.1
.35	160	511	31	52	83.8	90.8	89.0	75.5	94.3
.40	156	514	35	49	81.7	91.3	88.9	76.1	93.6
.45	154	523	37	40	80.6	92.9	89.8	79.4	93.4
.50	148	530	43	33	77.5	94.1	89.9	81.8	92.5
.55	137	532	54	31	71.7	94.5	88.7	81.5	90.8
.60	131	536	60	27	68.6	95.2	88.5	82.9	89.9
.65	126	541	65	22	66.0	96.1	88.5	85.1	89.3
.70	121	548	70	15	63.4	97.3	88.7	89.0	88.7
.75	114	552	77	11	59.7	98.0	88.3	91.2	87.8
.80	98	555	93	8	51.3	98.6	86.6	92.5	85.6
.85	89	558	102	5	46.6	99.1	85.8	94.7	84.5
.90	72	560	119	3	37.7	99.5	83.8	96.0	82.5

Abbreviations: PPV, positive predictive value; NPV, negative predictive value.

results of the present models (M1 and M2). In Figure 2, the results of model M1 are compared with the ROC of the Risk of Malignancy Index (RMI)[7] and the ROC of an old logistic regression model by Timmerman et al[9] applied to the test set cases with serum CA-125 results available (236 patients). The model M1 was significantly better than both the RMI ($P = .0038$) and the old logistic regression model ($P = .0187$). In the test set with time-averaged maximum velocity results available (220 patients), the model M1 performed significantly better (AUC = 0.95; SE = 0.014) than another old logistic regression model by Tailor et al[8] that included time-averaged maximum velocity as a predictive variable ($P = .0006$).

Table 6. Classification Table for the Test Data Set (n = 312)

	Correctly Classified		Incorrectly Classified						
Probability Level (*P*)	Malignant	Benign	Malignant As Benign	Benign As Malignant	Sensitivity (%)	Specificity (%)	Accuracy (%)	PPV (%)	NPV (%)
.01	74	55	1	182	98.7	23.2	41.3	28.9	98.2
.05	72	152	3	85	96.0	64.1	71.8	45.9	98.1
.10	70	179	5	58	93.3	75.5	79.8	54.7	97.3
.15	69	192	6	45	92.0	81.0	83.7	60.5	97.0
.20	66	200	9	37	88.0	84.4	85.3	64.1	95.7
.25	62	204	13	33	82.7	86.1	85.3	65.3	94.0
.30	62	208	13	29	82.7	87.8	86.5	68.1	94.1
.35	60	212	15	25	80.0	89.5	87.2	70.6	93.4
.40	60	219	15	18	80.0	92.4	89.4	76.9	93.6
.45	58	223	17	14	77.3	94.1	90.1	80.6	92.9
.50	58	226	17	11	77.3	95.4	91.0	84.1	93.0
.55	57	231	18	6	76.0	97.5	92.3	90.5	92.8
.60	55	234	20	3	73.3	98.7	92.6	94.8	92.1
.65	51	235	24	2	68.0	99.2	91.7	96.2	90.7
.70	48	236	27	1	64.0	99.6	91.0	98.0	89.7
.75	41	236	34	1	54.7	99.6	88.8	97.6	87.4
.80	36	237	39	0	48.0	100	87.5	100	85.9
.85	31	237	44	0	41.3	100	85.9	100	84.3
.90	23	237	52	0	30.7	100	83.3	100	82.0

Abbreviations: PPV, positive predictive value; NPV, negative predictive value.

Table 7. Comparison of the Performance of Different Models in the Test Data Set With CA-125 Results Available (n = 236)

	Area Under ROC Curve	SE	Cutoff	Sensitivity (%)	Specificity (%)
M1*	0.936	0.020	0.10	92.7	74.3
M2*	0.916	0.021	0.10	89.9	70.7
RMI*	0.870	0.028	100	78.3	79.6
Tailor et al[8]†	0.869	0.025	0.25	63.2	88.2
Timmerman et al[9]*	0.903	0.023	0.25	79.7	80.8

Abbreviations: CA-125, serum level of the tumor marker; ROC, receiver operating characteristic.

*Applied on the cases in the test set with CA-125 (236 cases).

†Applied only on the cases in the test set with time-averaged maximum velocity (220 cases).

Tumor Misclassification

There were 14 malignant masses incorrectly classified as benign when the model (M1) with a probability value of .10 was applied to the development data set. Of these 14 masses, there were 10 borderline malignant (of 40 borderline cases in the development set, ie, 25%), three primary invasive (of 121, ie, 2%), and one metastatic mass (of 30, ie, 3%). Under the same conditions, 130 benign masses were classified as malignant (of 563, ie, 23%).

DISCUSSION

This is the first report of a prospective, multicenter study to collect data for the development of a mathematical model to distinguish between malignant and benign overt adnexal masses before surgery. The study identified several defined indices that can be used to discriminate between malignant and benign masses. The developed models are based on a large number and variety of malignant and benign ovarian masses. We are optimistic that this development should provide a more robust model than previous versions. Furthermore, the multicenter aspect of this study implies that any equation derived from the study is more likely to be generally applicable to other populations. In this context, it is reassuring that when the model (M1) was tested on the data from each individual center, the overall test performance did not deteriorate.

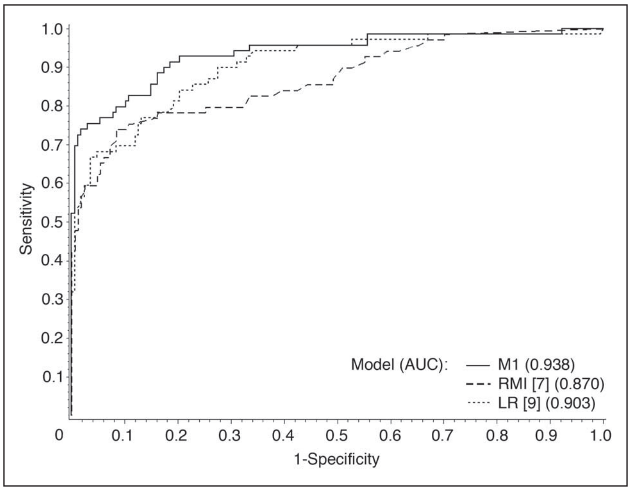

Fig 2. The receiver operating characteristic (ROC) curves of the logistic regression model (M1) and ROC of the Risk of Malignancy Index (RMI)[7] and ROC of an old logistic regression model (LR) by Timmerman et al[9] applied to the test set cases with serum CA-125 results available (n = 236). The areas under the curve (AUC) are 0.94, 0.87, and 0.90, respectively.

The results of the internal validation from this study also compare favorably with the results of previous attempts on smaller datasets to validate mathematical models prospectively.[12,13,14,22] Moreover, our model (M1) had significantly better test results on the test set cases than previously published models that had been developed on smaller datasets.[7,9] Our simpler model (M2) might be easier to use than the full model (M1). However, our full model included all significant variables, and the simpler model performed significantly less well than this full model on the test data set.

Although the area under the ROC curve reflects the overall test performance, the optimal probability level must be chosen to suit each individual clinical setting. There is always a compromise between the sensitivity (true-positive rate) and the number of false-positive test results. For example, by selecting a probability level of .10, the sensitivity for malignancy (model M1) was 93% and the specificity was 76%. In previous studies, we showed that highly experienced operators using subjective impression as the basis to define malignancy gave a sensitivity and specificity of 96% and 90%, respectively,[23] or 88% and 96%, respectively.[24] For less experienced operators, the corresponding values were 86% and 80%.[23] Therefore, our model could be helpful to less experienced operators.

Our study confirms the results of previous studies that simple unilocular cysts have only a small risk of being malignant.[25-30] We know that some benign pathology is relatively easy to characterize,[24,31] whereas late-stage cancers are usually quite obvious. Previous reports relating to subjective impression of ovarian masses suggest that some are difficult to classify even for the most experienced operators.[23,24] We have shown in our study that with sufficient numbers of patients and variables, a multimodal model can be built that will classify most tumors accurately before surgery. New models may be developed that concentrate on the characterization of more complex or difficult pathology.

The best model (M1) was retrospectively developed and then evaluated on a separate test set of patients. Although the internal validation of the model with our test data gave encouraging results, a true assessment can only be made by prospective evaluation on different study populations. In the near future, we will start the second phase of the

IOTA collaboration, where models M1 and M2 will be prospectively evaluated in new centers.

Appendix

Members of the IOTA Steering Committee: Dirk Timmerman, Lil Valentin, Thomas H. Bourne, William P. Collins, Sabine Van Huffel, and Ignace Vergote.

IOTA principal investigators (in alphabetical order): Jean-Pierre Bernard, Maurepas, France; Thomas H. Bourne, London, United Kingdom; Enrico Ferrazzi, Milan, Italy; Davor Jurkovic, London, United Kingdom; Fabrice Lécuru, Paris, France; Andrea Lissoni, Monza, Italy; Ulrike Metzger, Paris, France; Dario Paladini, Napels, Italy; Antonia Testa, Roma, Italy; Dirk Timmerman, Leuven, Belgium; Lil Valentin, Malmö, Sweden; Caroline Van Holsbeke, Leuven, Belgium; Sabine Van Huffel, Leuven, Belgium; Ignace Vergote, Leuven, Belgium; Gerardo Zanetta [deceased], Monza, Italy.

Authors' Disclosures of Potential Conflicts of Interest

The authors indicated no potential conflicts of interest.

REFERENCES

1. Vergote I, De Brabanter J, Fyles A, et al: Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. Lancet 357:176-182, 2001
2. 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 35:139-144, 1989
3. Sassone AM, Timor-Tritsch IE, Artner A, et al: Transvaginal sonographic characterisation of ovarian disease: Evaluation of a new scoring system to predict ovarian malignancy. Obstet Gynecol 78:70-76, 1991
4. Timor-Tritsch LE, Lerner JP, Monteagudo A, et al: Transvaginal ultrasonographic characterization of ovarian masses by means of color flow-directed Doppler measurements and a morphologic scoring system. Am J Obstet Gynecol 168:909-913, 1993
5. Lerner JP, Timor-Tritsch IE, Federman A, et al: Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring system. Am J Obstet Gynecol 170:81-85, 1994
6. Bourne T, Campbell S, Steer CV, et al: Transvaginal colour flow imaging: A possible new screening technique for ovarian cancer. BMJ 299:1367-1370, 1989
7. Jacobs I, Oram D, Fairbanks J, et al: A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. Br J Obstet Gynaecol 97:922-929, 1990
8. Tailor A, Jurkovic D, Bourne TH, et al: Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. Ultrasound Obstet Gynecol 10:41-47, 1997
9. Timmerman D, Bourne T, Tailor A, et al: A comparison of methods for the pre-operative discrimination between benign and malignant adnexal masses: The development of a new logistic regression model. Am J Obstet Gynecol 181:57-65, 1999
10. Timmerman D, Verrelst H, Bourne TH, et al: Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol 13:17-25, 1999
11. Tailor A, Jurkovic D, Bourne TH, et al: Sonographic prediction of malignancy in adnexal masses using an artificial neural network. Br J Obstet Gynaecol 106:21-30, 1999
12. Aslam N, Banerjee S, Carr JV, et al: Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer. Obstet Gynecol 96:75-80, 2000
13. Mol BWJ, Boll D, De Kanter M, et al.: Distinguishing the benign and malignant adnexal mass: An external validation of prognostic models. Gynecol Oncol 80:162-167, 2001
14. Valentin L, Hagen B, Tingulstad S, et al: Comparison of 'pattern recognition' and logistic regression models for discrimination between benign and malignant pelvic masses: A prospective cross validation. Ultrasound Obstet Gynecol 18:357-365, 2001
15. Aerts S, Antal P, Timmerman D, et al: Web based data collection for ovarian cancer: A case study. Proceedings of the 15th IEEE Symposium on Computer Based Medical Systems (CBMS), Maribor, Slovenia, 2002 (abstr)
16. Timmerman D, Valentin L, Bourne TH, et al: Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: A consensus opinion from the international ovarian tumor analysis (IOTA) group. Ultrasound Obstet Gynecol 16:500-505, 2000
17. Heintz APM, Odicino F, Maisonneuve P, et al: Carcinoma of the Ovary: 25th Annual Report on the Results of Treatment in Gynecological Cancer. Int J Gynaecol Obstet 83:S135-S137, 2003 (suppl 1)
18. Harrell FE Jr: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York, NY, Springer Series in Statistics, 2001
19. Hastie T, Tibshirani R: Generalized Additive Models: Monographs on Statistics and Applied Probability. Boca Ratan, FL, Chapman and Hall, 1999
20. Hosmer DW, Lemeshow S: Applied Logistic Regression: Wiley Series in Probability and Statistics. New York, NY, John Wiley & Sons, 2000
21. De Long ER, De Long DM, Clarke-Pearson DL: Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44:837-845, 1988
22. Timmerman D, Verrelst H, Collins WP, et al: Distinguishing the benign and malignant adnexal mass: An external validation of prognostic models. Gynecol Oncol 83:166-168, 2001
23. Timmerman D, Schwärzler P, Collins WP, et al: Subjective assessment of adnexal masses using ultrasonography: An analysis of interobserver variability and experience. Ultrasound Obstet Gynecol 13:11-16, 1999
24. Valentin L: Pattern recognition of pelvic masses by gray-scale ultrasound imaging: The contribution of Doppler ultrasound. Ultrasound Obstet Gynecol 14:338-347, 1999
25. Goldstein SR, Subramanyam B, Snyder J, et al: The postmenopausal cystic adnexal mass: The potential role of ultrasound in conservative management. Obstet Gynecol 73:8-10, 1988
26. Obwegeser R, Deutinger J, Bernascheck G: The risk of malignancy with an apparently simple adnexal cyst on ultrasound. Arch Gynecol Obstet 253:117-120, 1993
27. Kroon E, Andolf E: Diagnosis and follow-up of simple ovarian cysts detected by ultrasound in postmenopausal women. Obstet Gynecol 85: 211-214, 1995
28. Gerber B, Muller H, Kulz T, et al: Simple ovarian cysts in premenopausal patients. Int J Gynaecol Obstet 57:49-55, 1997
29. Bailey C, Ueland F, Land GL, et al: The malignant potential of small cystic ovarian tumors in women over 50 years of age. Gynecol Oncol 69:3-7, 1998
30. Ekerhovd E, Wienerroith H, Staudach A, et al: Preoperative assessment of unilocular adnexal cysts by transvaginal sonography: A comparison between sonographic morphological imaging and histopathologic diagnosis. Am J Obstet Gynecol 184:48-54, 2001
31. Jermy K, Luise C, Bourne T: The characterization of common ovarian cysts in premenopausal women. Ultrasound Obstet Gynecol 17: 140-144, 2001