Prospective Internal Validation of Mathematical Models to Predict Malignancy in Adnexal Masses: Results from the International Ovarian Tumor Analysis Study

Caroline Van Holsbeke,[1,3] Ben Van Calster,[2] Antonia C. Testa,[4] Ekaterini Domali,[1] Chuan Lu,[5] Sabine Van Huffel,[2] Lil Valentin,[6] and Dirk Timmerman[1]

Abstract

Purpose: To prospectively test the mathematical models for calculation of the risk of malignancy in adnexal masses that were developed on the International Ovarian Tumor Analysis (IOTA) phase 1 data set on a new data set and to compare their performance with that of pattern recognition, our standard method.

Methods: Three IOTA centers included 507 new patients who all underwent a transvaginal ultrasound using the standardized IOTA protocol. The outcome measure was the histologic classification of excised tissue. The diagnostic performance of 11 mathematical models that had been developed on the phase 1 data set and of pattern recognition was expressed as area under the receiver operating characteristic curve (AUC) and as sensitivity and specificity when using the cutoffs recommended in the studies where the models had been created. For pattern recognition, an AUC was made based on level of diagnostic confidence.

Results: All IOTA models performed very well and quite similarly, with sensitivity and specificity ranging between 92% and 96% and 74% and 84%, respectively, and AUCs between 0.945 and 0.950. A least squares support vector machine with linear kernel and a logistic regression model had the largest AUCs. For pattern recognition, the AUC was 0.963, sensitivity was 90.2%, and specificity was 92.9%.

Conclusion: This internal validation of mathematical models to estimate the malignancy risk in adnexal tumors shows that the IOTA models had a diagnostic performance similar to that in the original data set. Pattern recognition used by an expert sonologist remains the best method, although the difference in performance between the best mathematical model is not large.

Authors' Affiliations: [1]Department of Obstetrics and Gynecology, University Hospitals Leuven; [2]Department of Electrical Engineering, K.U. Leuven, Leuven, Belgium; [3]Department of Obstetrics and Gynecology, Ziekenhuis Oost-Limburg, Genk, Belgium; [4]Instituto di Clinica Ostetrica e Ginecologica, Universita Cattolica del Sacro Cuore, Rome, Italy; [5]Department of Computer Science, University of Wales, Aberystwyth, Wales; and [6]Department of Obstetrics and Gynecology, Malmö University Hospital, Lund University, Malmö, Sweden

Received 1/14/08; revised 10/2/08; accepted 10/2/08.

Grant support: Research Council of the Katholieke Universiteit Leuven (GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering OPTEC), Belgian Federal Science Policy Office IUAP P6/04 ("Dynamical Systems, Control and Optimization," 2007-2011), EU: BIOPATTERN (FP6-2002-IST 508803), ETUMOUR (FP6-2002-LIFESCIHEALTH 503094), Healthagents (IST-2004-27214), Swedish Medical Research Council (grants K2001-72X-11605-06A, K2002-72X-11605-07B, K2004-73X-11605-09A, and K2006-73X-11605-11-3), Malmö University Hospital, Allmänna Sjukhusets i Malmö Stiftelse för bekämpande av cancer (Malmö General Hospital Foundation for Fighting Against Cancer), and ALF-medel and Landstingsfinansierad regional forskning (two Swedish governmental grants).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Dirk Timmerman, Department of Obstetrics and Gynecology, University Hospitals Leuven, K.U. Leuven, Herestraat 49, B-3000 Leuven, Belgium. Phone: 32-16-344201; Fax: 32-16-344205; E-mail: dirk.timmerman@uz.kuleuven.ac.be.

doi:10.1158/1078-0432.CCR-08-0113

Correct preoperative discrimination between benign and malignant adnexal masses is important, because the preoperative diagnosis will determine the treatment of the patient. Incorrect diagnosis is likely to result in incorrect treatment and this may worsen the prognosis. Because benign and malignant adnexal masses show different ultrasound morphology, a transvaginal ultrasound examination can be used to discriminate preoperatively between benign and malignant masses (1, 2), one of the best ultrasound methods being pattern recognition, that is, subjective evaluation of gray-scale and Doppler ultrasound findings by the ultrasound examiner (3, 4). Several scoring systems and mathematical models using ultrasound variables have been developed for the preoperative prediction of probability of malignancy. However, before these models can be used in daily practice, they should be tested prospectively in new populations. Some of these models have been tested prospectively on small data sets with disappointing results (5–7).

We have prospectively tested 17 scoring systems and mathematical models for calculation of malignancy risk in adnexal masses on a large data set collected by the International Ovarian Tumor Analysis (IOTA) collaborative group, the IOTA phase 1 data set (8). Unfortunately, most models and scoring systems performed worse than they had done in the studies

Translational Relevance

For clinicians, ultrasound pattern recognition is the standard method for the preoperative prediction of malignancy in an adnexal mass. In the hands of experienced ultrasound examiners, this strategy results in a sensitivity of 96% and a specificity of 90% (1). Less experienced examiners using pattern recognition will probably not achieve such high sensitivity and specificity. Therefore, the development of mathematical models to predict malignancy in adnexal masses is worthwhile. Because the prospective testing of most previously published models was disappointing, most likely because the models had been developed in a single small center and because neither examination technique nor terms to describe the ultrasound findings had been standardized, we developed 11 new models in a multicenter study (the IOTA study) including 1,066 patients scanned using the same standardized ultrasound protocol. In phase 1 of the IOTA study, all models proved to perform very well with AUC of >0.92. This area is larger than that of the risk of malignancy index, which is often regarded as a standard test (2–6). Internal validation is the next step in the evaluation process of new diagnostic tests. This article describes the internal validation of our new models. The AUCs of all models were similar to those in the IOTA study phase 1 and to that of pattern recognition. These results are promising, but external validation of the models in the hands of less experienced examiners remains to be done.

where they had been created (8). However, the primary aim of the IOTA phase 1 study was not to prospectively test previously developed methods for discriminating between benign and malignant adnexal masses but to prospectively collect clinical information and standardized data from ultrasound examinations in a large number of patients with adnexal tumors to create new scoring systems and mathematical models to distinguish benign from malignant adnexal tumors (9). More than 50 variables of 1,066 patients with 1,233 masses were analyzed. Most variables were gray-scale and color Doppler variables, but clinical and demographic variables were recorded as well. All patients were scanned by an expert sonologist of the IOTA group in one of the nine European centers following a strict protocol that has been published previously (10). For the development of all models, the database was divided into a training set of 754 patients and an independent test set of 312 patients. All models performed well when tested on the test set of patients (n = 312). Their area under the receiver operating characteristics (ROC) curve (AUC) ranged from 0.93 to 0.95 (9, 11, 12).

The aim of the present study was to prospectively test the mathematical models developed on the IOTA phase 1 data set on a new data set and to compare their performance with that of pattern recognition, which is assumed to be the standard method.

Materials and Methods

New prospective data set. Patients were recruited from three ultrasound centers: University Hospitals Leuven, Malmö University Hospital, and Instituto di Clinica Ostetrica e Ginecologica. These centers had contributed two thirds of the cases to the IOTA phase 1 data set (25%, 30%, and 12%, respectively). Consecutive patients meeting the inclusion criteria of the IOTA phase 1 study (9) underwent ultrasound examination by the same examiners as in the IOTA phase 1 study following the same IOTA study protocol and using the same or similar ultrasound systems as in the IOTA phase 1 study. Patient recruitment started immediately after completion of the IOTA phase 1 study (June 2002) and finished in December 2005. The IOTA study protocol and the IOTA terms and definitions have been described in detail elsewhere (9, 10). In accordance with the IOTA phase 1 study protocol, the ultrasound examiner was obliged to give his/her subjective impression in two ways: (*a*) classification of each mass as benign or malignant based on subjective evaluation of ultrasound findings (pattern recognition) and (*b*) expressing his/her level of confidence as follows: benign, probably benign, uncertain, probably malignant, or malignant. The category "uncertain" was split into two subcategories: uncertain but first classified as benign and uncertain but first classified as malignant. The criteria described in ref. 13 were used for pattern recognition.

Pattern recognition and the following mathematical models developed on the IOTA phase 1 data set were tested on the newly collected study population: two logistic regression models [LR1 (12 variables) and LR2 (6 variables); ref. 9], three least squares support vector machine (LS-SVM) models (LS-SVM lin, LS-SVM rbf, and LS-SVM add rbf; ref. 11), three relevance vector machine (RVM) models (RVM lin, RVM rbf, and RVM add rbf; ref. 3), and three neural networks (Bay MLP 11-2a, Bay MLP 11-2b, and Bay Perc 11; ref. 12).

LR1 was developed using statistical state-of-the-art logistic regression modeling, with due attention to multicollinearity, possible interactions, and linearity in the logit assumption. It held back 12 independent variables to predict malignancy. These variables were selected using both automated and manual selection procedures. LR1 gave an AUC of 0.942 on the test set (n = 312). Sensitivity and specificity were 93% and 76%, respectively, using a cutoff of 0.10. LR1 was the main logistic regression model. A reduced model (LR2) using only 6 variables was developed as well (9).

Kernel methods were also applied for binary classification purposes in refs. 12, 14–16. Methods considered are Bayesian LS-SVM (17) and RVM (17) with linear, radial basis function (RBF), and additive RBF kernels. LS-SVMs and RVMs are mathematically advanced methods that are flexible in creating a nonlinear separation between two classes. Mathematically, these methods make a linear separation after applying a kernel function to the data. Relative to the original data, the use of a kernel function may introduce nonlinearity in the separation between both classes. RVMs work with any function (not necessarily with a positive semidefinite kernel function), such that RVMs are not true kernel-based methods, whereas LS-SVMs are. Using a linear kernel, the separation between classes will still be linear with respect to the original data. A typical kernel function to obtain nonlinear separation is the RBF. The additive RBF kernel (18) is an extension of the RBF kernel to get more insight in how the model works. More explanation on LS-SVMs and RVMs can be found in refs. 11, 17, 18. More information about the Bayesian approach to LS-SVMs is given in ref. 11. Variable selection was done using forward and backward selection in a Bayesian LS-SVM framework. The variable selection analyses resulted in 12 variables being included in the mathematical models (Table 1). When these models were tested on the test set of IOTA phase 1, all models had similar performance, with test set AUCs ranging from 0.94 to 0.95, accuracy from 83% to 86%, sensitivity from 91% to 93%, and specificity from 81% to 84%.

Next to the kernel methods, a Bayesian multilayer perceptron using the evidence procedure (19, 20) was applied as another type of model that can incorporate nonlinearity (14). Variable selection was done using Automatic Relevance Determination and model selection using cross-validation. A linear model using Bayesian perceptron was also developed. The best Bayesian multilayer perceptron included 11 variables with two hidden neurons (Bay MLP 11-2a). All three neural networks used different sets of variables; the models included 6 to 12 of the variables, which are listed in Table 1.

Table 1. Description of all 11 models developed on the IOTA phase 1 data set

	Type of model	Cutoff	Variables
LR1 (9)	Logistic regression	0.10	(1) personal history of ovarian cancer, (2) previous use of hormonal therapy, (3) age, (4) maximum diameter of the lesion, (5) pain, (6) ascites, (7) blood flow within papillary projection, (8) presence of solid tumor, (9) maximal diameter of solid component (bounded at 50 mm), (10) irregular internal cyst walls, (11) acoustic shadows, and (12) color score of intratumoral blood flow
LR2 (9)	Logistic regression	0.10	(1) age, (2) ascites, (3) blood flow within a solid papillary projection, (4) maximal diameter of the solid component (bounded at 50 mm), (5) irregular internal cyst walls, and (6) acoustic shadows
Bay MLP 11-2a (12)	Artificial neural network	0.15	(1) age, (2) hormonal therapy, (3) ascites, (4) maximum diameter of the lesion, (5) irregular internal cyst wall, (6) color score, (7) blood flow within papillary projection, (8) number of papillary projection, (9) maximum diameter of solid component, (10) multilocular-solid tumor, and (11) solid tumor
Bay MLP 11-2b (12)	Artificial neural network	0.15	(1) age, (2) personal history of ovarian cancer, (3) pelvic pain during examination, (4) ascites, (5) maximum diameter of the lesion, (6) irregular internal cyst wall, (7) blood flow within papillary projection, (8) acoustic shadowing, (9) maximum diameter of solid component, (10) unilocular tumor, and (11) solid tumor
Bay Perc 11 (12)	Artificial neural network	0.15	(1) age, (2) personal history of ovarian cancer, (3) ascites, (4) maximum diameter of the lesion, (5) irregular internal cyst wall, (6) color score, (7) blood flow within papillary projection, (8) acoustic shadowing, (9) maximum diameter of solid component, (10) unilocular tumor, and (11) solid tumor
LS-SVM lin (11)	Support vector machine	0.15	(1) maximum diameter of the solid component, (2) maximum diameter of the ovary, (3) age, (4) color score 4, (5) presence of a multilocular-solid lesion, (6) ascites, (7) personal history of ovarian cancer, (8) blood flow within papillary projection, (9) acoustic shadows, (10) previous use of hormonal therapy, (11) irregular internal cyst walls, and (12) whether the tumor is suspected to be of ovarian origin
LS-SVM rbf (11)	Support vector machine	0.12	(1) maximum diameter of the solid component, (2) maximum diameter of the ovary, (3) age, (4) color score 4, (5) presence of a multilocular-solid lesion, (6) ascites, (7) personal history of ovarian cancer, (8) blood flow within a papillary projection, (9) acoustic shadows, (10) previous use of hormonal therapy, (11) irregular internal cyst walls, and (12) whether the tumor is suspected to be of ovarian origin
LS-SVM add rbf (11)	Support vector machine	0.12	(1) maximum diameter of the solid component, (2) maximum diameter of the ovary, (3) age, (4) color score 4, (5) presence of a multilocular-solid lesion, (6) ascites, (7) personal history of ovarian cancer, (8) blood flow within a papillary projection, (9) acoustic shadows, (10) previous use of hormonal therapy, (11) irregular internal cyst walls, and (12) whether the tumor is suspected to be of ovarian origin
RVM lin (11)	RVM	0.20	(1) maximum diameter of the solid component, (2) maximum diameter of the ovary, (3) age, (4) color score 4, (5) presence of a multilocular-solid lesion, (6) ascites, (7) personal history of ovarian cancer, (8) blood flow within a papillary projection, (9) acoustic shadows, (10) previous use of hormonal therapy, (11) irregular internal cyst walls, and (12) whether the tumor is suspected to be of ovarian origin
RVM rbf (11)	RVM	0.15	(1) maximum diameter of the solid component, (2) maximum diameter of the ovary, (3) age, (4) color score 4, (5) presence of a multilocular-solid lesion, (6) ascites, (7) personal history of ovarian cancer, (8) blood flow within a papillary projection, (9) acoustic shadows, (10) previous use of hormonal therapy, (11) irregular internal cyst walls, and (12) whether the tumor is suspected to be of ovarian origin
RVM add rbf (11)	RVM	0.15	(1) maximum diameter of the solid component, (2) maximum diameter of the ovary, (3) age, (4) color score 4, (5) presence of a multilocular-solid lesion, (6) ascites, (7) personal history of ovarian cancer, (8) blood flow within a papillary projection, (9) acoustic shadows, (10) previous use of hormonal therapy, (11) irregular internal cyst walls, and (12) whether the tumor is suspected to be of ovarian origin

Statistical analysis. The diagnostic performance of each model was expressed as the AUC (21) and the partial AUC (22). The partial AUC is the area under that part of the ROC curve, which is defined by the lowest acceptable specificity. Because we believe that the lowest acceptable specificity when predicting malignancy in ovarian tumors is 75%, we computed the partial AUC (0.75), that is, the area under that part of the ROC curve where the specificity is ≥75% (the most left part of the ROC curve; ref. 21). Using the six levels of diagnostic confidence as different cutoffs, a ROC curve for pattern recognition could be constructed as well. The diagnostic performance of the models was also expressed as sensitivity, specificity, positive and negative predictive value, and positive and negative likelihood ratio when using the risk cutoff recommended in the original article describing the model. However, sensitivity and specificity of a model depend on the cutoff chosen, whereas the AUC reflects overall test performance. Therefore, we considered the AUC to be the most important measure of diagnostic performance (20). Because the statistical comparison of all 11 models would yield 55 different *P* values, we preferred to compare the performance of the models by using the ranking method based on the work of Pepe et al. (23). In this method, all models are ranked with regard to a chosen criterion. For our analysis, the main criterion was the AUC. Because the ranking is influenced by sampling variability, the probability that a method is ranked among the κ best methods, Pm(κ), is computed using 1,000 bootstrap samples (23). A bootstrap sample is a new data set generated from the original data set, where both have the same sample size. The bootstrap sample is constructed by randomly selecting patients from the original data set. Patients are selected "with replacement," meaning that a patient can be selected more than once or not at all for a given bootstrap sample (25). For each bootstrap sample, the AUCs were computed and the models were ranked according to their AUC. This allowed us to compute how many times each model was ranked as the best, Pm(1), among the best three, Pm(3), and among the best five, Pm(5). The mean rank over all bootstrap samples was also computed.

To determine differences between benign and malignant tumors in numerical ultrasound and clinical variables, the θ index for effect size [with its 95% confidence interval (95% CI)] was used as the main indicator (24). The θ index can take on any value between 0.5 and 1 and can be interpreted as the degree of overlap between the benign and the malignant groups. A θ value of 1 means no overlap and a θ value of 0.5 means maximal overlap. The θ index is mathematically identical to the AUC, but a ROC curve analysis has different objectives (25). Nonetheless, this relationship may help nonstatisticians to interpret θ. For dichotomous variables, the difference between percentages was computed, together with a 95% CI using method 10 as described by Newcombe (26).

To compare levels of sensitivity or specificity between different approaches, we computed the difference in sensitivity or specificity and constructed a 95% CI for paired proportions using method 10 as described by Newcombe (27).

Results

The prospectively collected study population consists of 507 patients with complete information for all 17 variables used in the models tested. Of these 507 patients, 287 (57%) were examined in Leuven, Belgium, 96 (19%) in Rome, Italy, and 124 (24%) in Malmö, Sweden. The malignancy rate in the whole study population was 28% (143 of 507), with a malignancy rate of 30% (85 of 287) in Leuven, 33% (32 of 96) in Rome, and 21% (26 of 124) in Malmö. Histologic diagnoses are shown in Table 2. Tables 3 and 4 present clinical data and ultrasound findings in benign and malignant tumors separately for each of the three participating centers.

Diagnostic performance of the models tested

The performance of the models tested is shown in Table 5 together with the results of pattern recognition. The models performed very well and quite similarly, with AUCs ranging from 0.945 to 0.950, sensitivity from 91.6% to 95.1%, and specificity from 73.9% to 83.8%. A LS-SVM with linear kernel and the logistic regression model LR1 had the largest AUC (0.950). The largest difference in AUC between any two models was only 0.005 (Table 5).

The logistic regression model LR2 with 6 variables had the highest partial AUC (0.75), that is, 0.208. By using the ranking method for AUCs of Pepe et al. (23), the LS-SVM with linear kernel and with rbf kernel and the two logistic regression models

Table 2. Tumor type and histology in each of the three participating centers

Histology	All (*n* = 507)	Leuven (*n* = 287)	Rome (*n* = 96)	Malmö (*n* = 124)
Benign	364 (72.0)	202 (70.7)	64 (66.7)	98 (79.0)
Endometrioma	101 (19.9)	62 (21.6)	16 (16.7)	23 (18.6)
Teratoma	55 (10.9)	30 (14.5)	12 (12.5)	13 (10.5)
Simple cyst	48 (9.5)	30 (14.5)	6 (6.3)	12 (9.7)
Functional cyst	11 (2.2)	4 (1.4)	3 (3.1)	4 (3.2)
Hydrosalpinx	16 (3.2)	8 (2.8)	1 (1.0)	7 (5.7)
Peritoneal pseudocyst	5 (1.0)	4 (1.4)	1 (1.0)	0 (0.0)
Abscess	4 (0.8)	1 (0.3)	3 (3.1)	0 (0.0)
Fibroma	27 (5.3)	15 (5.2)	9 (9.4)	3 (2.4)
Serous cystadenoma	48 (9.5)	24 (8.4)	6 (6.3)	18 (14.5)
Mucinous cystadenoma	38 (7.5)	15 (5.2)	7 (7.3)	16 (12.9)
Rare benign	11 (2.2)	9 (3.1)	0 (0.0)	2 (1.6)
Primary invasive	103 (20.1)	62 (21.3)	24 (25.0)	17 (13.7)
Primary invasive stage I	24 (4.7)	14 (4.9)	7 (7.3)	3 (2.4)
Primary invasive stage II	5 (1.0)	2 (0.7)	0 (0.0)	3 (2.4)
Primary invasive stage III	59 (11.6)	37 (12.9)	16 (16.7)	6 (4.8)
Primary invasive stage IV	11 (2.2)	9 (3.1)	0 (0.0)	2 (1.6)
Rare primary invasive	4 (0.8)	0 (0.0)	1 (1.0)	3 (2.4)
Borderline	20 (3.9)	12 (4.2)	2 (2.1)	6 (4.8)
Metastatic	20 (3.9)	11 (3.8)	6 (6.3)	3 (2.4)

NOTE: All values expressed as *n* (%).

Table 3. Clinical data and ultrasound findings (continuous and categorical data) in benign and malignant tumors presented separately for each of the three participating centers

Leuven new (n = 287)	Benign		Malignant		θ (95% CI)
Variable	***n***	**Median**	***n***	**Median**	
Age (y)	202	40	85	58	0.77 (0.70-0.82)
Max diam lesion, mm	202	52	85	89	0.70 (0.63-0.76)
No. papillations	37	2	49	>3	0.74 (0.62-0.83)
Max diam solid, mm	82	26.5	82	56.5	0.71 (0.62-0.78)
Color score, blood flow	202	2	85	3	0.87 (0.81-0.91)
CA 125 (units/mL)	109	17	81	296	0.88 (0.82-0.92)
Rome new (n = 96)	**Benign**		**Malignant**		**θ (95% CI)**
Variable	***n***	**Median**	***n***	**Median**	
Age (y)	64	45	32	52	0.64 (0.51-0.74)
Max diam lesion, mm	64	78	32	101	0.64 (0.51-0.74)
No. papillations	14	2	9	>3	0.65 (0.41-0.83)
Max diam solid, mm	25	49	30	64	0.67 (0.52-0.79)
Color score, blood flow	64	1	32	3	0.82 (0.71-0.89)
CA 125 (units/mL)	20	38.5	22	220	0.75 (0.57-0.86)
Malmö new (n = 124)	**Benign**		**Malignant**		**θ (95% CI)**
Variable	***n***	**Median**	***n***	**Median**	
Age (y)	98	42	26	55.5	0.74 (0.62-0.83)
Max diam lesion, mm	98	68.5	26	116.5	0.77 (0.65-0.85)
No. papillations	19	1	11	>3	0.73 (0.51-0.87)
Max diam solid, mm	37	20	26	62	0.85 (0.73-0.92)
Color score, blood flow	98	3	26	3	0.68 (0.56-0.78)
CA 125 (units/mL)	88	18	25	251	0.86 (0.75-0.92)

Abbreviations: Max diam lesion, largest diameter of the lesion; Max diam solid, largest diameter of the largest solid component of the tumor.

were ranked among the five best performing models. However, Pm(1) ranked LS-SVM with rbf on the seventh place (Table 5).

Using the subjective classification as benign or malignant, sensitivity was 90.2% and specificity was 92.9%. This corresponds to a positive LR of 12.7 and a negative LR of 0.11 (Table 5; Fig. 1). After having expressed the level of diagnostic confidence with which the diagnosis of benignity or malignancy was made by the expert sonologists, they seemed to be uncertain about their diagnosis in 8% (39 of 507) of the cases and completely confident about the benign or malignant character of the mass in 46% (226 of 507) of the cases. The use of the different levels of confidence resulted in an AUC for pattern recognition of 0.963.

Finally, Table 6 and Fig. 1 compare the performance of the best IOTA model (LS-SVM lin; ref. 11) with the assumed standard method, pattern recognition. The AUC for LS-SVM lin was 0.950 versus 0.963 for pattern recognition (difference = -0.013; 95% CI, -0.034 to 0.005). Using the cutoff that was reported in the original publication, sensitivity was 91.6% for LS-SVM lin versus 90.2% for pattern recognition (difference = 1.4%; 95% CI, -4.2% to 7.1%) and specificity was 83.8% versus 92.9% for pattern recognition (difference = -9.1%; 95% CI, -13.1% to -5.2%).

Results per center

Leuven. All IOTA models performed very well when tested on the data collected in Leuven. The AUC values ranged from 0.947 to 0.958. A Bayesian neural network had the largest AUC (0.958). The two logistic regression models also had large AUCs (0.956 and 0.957; Table 5).

Rome. The performance of the IOTA models was moderate to good when tested on the data collected in Rome. The AUCs ranged from 0.878 to 0.905 with the largest AUC for a Bayesian perceptron network (Table 5).

Malmö. All IOTA models performed very well on the data collected in Malmö. The AUCs ranged from 0.975 to 0.992. The logistic regression model LR1 had the largest AUC (Table 5).

Discussion

Pattern recognition by an experienced sonologist is an excellent method for discriminating between benign and malignant adnexal masses and should probably be regarded as the standard method for preoperative classification of adnexal masses (4, 13). However, the ability to discriminate between benign and malignant adnexal masses using pattern recognition increases with increasing experience (1), and in daily clinical practice, it is impossible to ask an expert's opinion on every adnexal mass. The aim of developing mathematical models is to improve the less experienced examiner's ability to discriminate between benign and malignant adnexal masses so that it approaches that of an expert ultrasound examiner.

In the United Kingdom, the Royal College of Obstetricians and Gynaecologists recommends the use of the risk of malignancy index (28, 29), but in ref. 9, we compared the performance of the two IOTA logistic regression models with the risk of malignancy index on a test set of 236 patients and showed that the risk of malignancy index performed

significantly worse: AUC 0.936 and 0.916 for the logistic regression models versus 0.870 for risk of malignancy index (P = 0.0038; ref. 9). The strategy followed by the American College of Gynecologists is the use of guidelines based on demographic and ultrasound variables (CA 125 level: >35 units/mL for postmenopausal patients and >200 units/mL for premenopausal patients, evidence of ascites on ultrasound or computed tomography, a nodular or fixed pelvic mass, evidence of abdominal or distant metastases on computed tomography, and family history of at least one first-degree relative with ovarian or breast cancer) to classify an adnexal mass as benign or malignant and to select which patient should be referred to a tertiary center with a gynecologic oncology department (30).

When the performance of these guidelines was tested on a data set of 837 patients, the guidelines reached a sensitivity of 79% and 93% for the premenopausal and postmenopausal groups, respectively, and a specificity of 70% and 60%. After a revision of these guidelines that decreased the cutoff level of serum CA 125 in the premenopausal group (from 200 to 67 units/mL), sensitivity was 75% and 91% for premenopausal and postmenopausal patients, respectively, and specificity 91% and 76%.

Because the performance of these "gold Royal College of Obstetricians and Gynaecologists and American College of Gynecologists standards" were not optimal and the performance of the IOTA models on a test set of 236 patients was very good, we believed that the IOTA models deserved internal and external validation. Unfortunately, we cannot compare the performance of the IOTA models with that of the Royal College of Obstetricians and Gynaecologists or American College of Gynecologists guidelines because we have not prospectively collected information on physical examination or computed tomography findings concerning the presence of metastases.

In this internal validation, the IOTA models performed as well or better than they had done in the original studies where they had been created.

Table 4. Clinical data and ultrasound findings (binary data) in benign and malignant tumors presented separately for each of the three participating centers

Leuven new (n = 287)	Benign		Malignant		Difference in % (95% CI)
Variable	***n***	**% Yes**	***n***	**% Yes**	
Personal history of ovarian cancer	202	2 (1.0)	85	4 (4.7)	-3.7 (-10.5 to 0.1)
Hormonal therapy	202	38 (18.8)	85	17 (20.0)	-1.2 (-12.0 to 8.1)
Pelvic pain	202	71 (35.1)	85	20 (23.5)	11.6 (-0.2 to 21.9)
Ascites	202	3 (1.5)	85	47 (55.3)	-53.8 (-64.0 to -42.9)
Pap blood flow	37	7 (18.9)	49	43 (87.8)	-68.9 (-80.3 to -49.4)
Internal wall, irregular	202	93 (46.0)	85	62 (72.9)	-26.9 (-37.6 to -14.5)
Acoustic shadows	202	40 (19.8)	85	1 (1.2)	18.6 (11.5-24.7)
Unilocular tumor	202	88 (43.6)	85	1 (1.2)	42.4 (34.0-49.4)
Multilocular-solid tumor	202	25 (12.4)	85	38 (44.7)	-32.3 (-43.6 to -20.9)
Entirely solid tumor	202	30 (14.9)	85	29 (34.1)	19.2 (-30.7 to -8.5)
Rome new (n = 96)	**Benign**		**Malignant**		**Difference in % (95% CI)**
Variable	***n***	**% Yes**	***n***	**% Yes**	
Personal history of ovarian cancer	64	0 (0.0)	32	0 (0.0)	0.0 (-10.7 to 5.7)
Hormonal therapy	64	9 (14.1)	32	0 (0.0)	14.1 (1.5-24.6)
Pelvic pain	64	20 (31.3)	32	12 (37.5)	-6.2 (-26.2 to 12.7)
Ascites	64	1 (1.6)	32	5 (15.6)	-14.0 (-30.2 to -3.0)
Pap blood flow	14	4 (28.6)	9	7 (77.8)	-49.2 (-72.4 to -7.5)
Internal wall, irregular	64	26 (40.6)	32	25 (78.1)	-37.5 (-53.1 to -16.7)
Acoustic shadows	64	9 (14.1)	32	1 (3.1)	10.9 (-3.3 to 21.8)
Unilocular tumor	64	26 (40.6)	32	1 (3.1)	37.5 (20.7-50.0)
Multilocular-solid tumor	64	7 (10.9)	32	9 (28.1)	-17.2 (-35.3 to -1.2)
Entirely solid tumor	64	7 (10.9)	32	16 (50.0)	-39.1 (-56.3 to -19.9)
Malmö new (n = 124)	**Benign**		**Malignant**		**Difference in % (95% CI)**
Variable	***n***	**% Yes**	***n***	**% yes**	
Personal history of ovarian cancer	98	1 (1.0)	26	0 (0.0)	1.0 (-11.9 to 5.6)
Hormonal therapy	98	30 (30.6)	26	3 (11.5)	19.1 (-0.2 to 31.4)
Pelvic pain	98	8 (8.2)	26	1 (3.8)	4.4 (-11.2 to 12.1)
Ascites	98	0 (0.0)	26	2 (7.7)	-7.7 (-24.1 to -1.0)
Pap blood flow	19	8 (42.1)	11	11 (100)	-57.9 (-76.9 to -24.2)
Internal wall, irregular	98	29 (29.6)	26	21 (80.8)	-51.2 (-64.6 to -30.2)
Acoustic shadows	98	16 (16.3)	26	0 (0.0)	16.3 (2.1-24.9)
Unilocular tumor	98	28 (28.6)	26	0 (0.0)	28.6 (13.4-38.2)
Multilocular-solid tumor	98	25 (25.5)	26	12 (46.2)	-20.7 (-40.5 to -0.9)
Entirely solid tumor	98	4 (4.1)	26	10 (38.5)	-34.4 (-53.6 to -17.3)

Abbreviations: Pelvic pain, pain during the ultrasound examination; Pap blood flow, flow present within papillary projection.

Table 5. Performance of the models developed on the IOTA phase 1 data set when tested on the test set of the IOTA phase 1 data set and when tested prospectively on the current study population

Model	AUC performance			Cutoff performance				Ranking analysis		
	AUC IOTA phase 1 test set (95% CI), *n* = 312	AUC all (95% CI), *N* = 507	AUC Leuven-Rome-Malmö, *n* = 287-96-124	Sensitivity	Specificity	LR+	LR-	Mean rank	Pm(1)	Pm(3)
LS-SVM lin (11)	0.946 (0.911-0.973)	0.950 (0.931-0.967)	0.952-0.889-0.986	91.6	83.8	5.65	0.10	3.7	**0.148**	0.485
LR 1 (9)	0.942 (0.905-0.972)	0.950 (0.929-0.967)	0.956-0.903-0.992	95.1	74.2	3.68	0.07	4.4	**0.145**	0.472
LR 2 (9)	0.922 (0.882-0.954)	0.950 (0.928-0.967)	0.957-0.884-0.983	95.1	73.9	3.64	0.07	4.8	**0.209**	0.423
LS-SVM rbf (11)	0.945 (0.909-0.971)	0.949 (0.930-0.966)	0.952-0.891-0.981	91.6	83.2	5.47	0.10	4.6	0.048	**0.339**
Bay MLP 11-2a (12)	0.942 (0.900-0.974)	0.948 (0.928-0.964)	0.949-0.912-0.985	93.7	76.9	4.06	0.08	5.9	**0.150**	0.303
LS-SVM add rbf (11)	0.943 (0.905-0.970)	0.948 (0.929-0.963)	0.954-0.885-0.980	93.7	81.9	5.17	0.08	6.3	0.082	0.204
RVM lin (11)	0.949 (0.913-0.975)	0.948 (0.926-0.965)	0.949-0.892-0.984	91.6	82.1	5.13	0.10	6.6	0.024	0.137
Bay MLP 11-2b (12)	0.933 (0.892-0.964)	0.947 (0.926-0.966)	0.958-0.887-0.975	95.1	78.8	4.50	0.06	6.4	0.118	0.293
RVM add rbf (11)	0.946 (0.910-0.973)	0.947 (0.927-0.964)	0.954-0.878-0.980	93.7	81.3	5.02	0.08	7.0	0.026	0.128
Bay Perc 11 (12)	0.947 (0.908-0.974)	0.946 (0.923-0.963)	0.953-0.905-0.982	95.8	78.3	4.41	0.05	7.7	0.045	0.186
RVM rbf (11)	0.946 (0.910-0.972)	0.945 (0.923-0.964)	0.947-0.892-0.981	92.3	80.2	4.67	0.10	8.6	0.001	0.022
Pattern recognition		0.963		90.2	92.9	12.7	0.11			

NOTE: Because of the ranking analysis, the models are ranked from high to low AUC for all three centers combined.
Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio. Ranking analysis: Pm(1) gives the chance that the model is the best performing model; Pm(3) gives the chance that the model is one of the three best performing models; the four top models in each ranking category are marked in bold.

The robustness of the IOTA models may be explained by them having been developed on a large data set where data had been collected from many centers and using a standardized ultrasound protocol and standardized terms and definitions. On the other hand, the models were validated in the three centers that had provided most of the data to the IOTA phase 1 study, which means that our study presents the results of internal validation only. The models performed better in the centers in Leuven and Malmö than in Rome. This may be explained by differences in pathology between centers (Table 2). The center in Rome included more primary invasive tumors and metastatic tumors but fewer borderline tumors than the other centers, and it included more fibromas and abscesses. Fibromas and abscesses are often misclassified by mathematical models because of their solid character and high vascularity. Differences in performance between centers may also be explained by differences between ultrasound examiners in their evaluation of certain ultrasound features (e.g., in the evaluation of color score, which is highly subjective) and by differences in ultrasound equipment (the Doppler sensitivity of the ultrasound system being important when assessing the color score).

Pattern recognition in the hands of an experienced ultrasound examiner is a very good method for discriminating between benign and malignant adnexal tumors. The mathematical models seem to have slightly higher sensitivity and lower specificity than pattern recognition, but this is dependent on the cutoff chosen for the mathematical model. In this study, the mathematical models performed nearly as well as pattern recognition when comparing the ROC curves. The models now need to undergo external validation, that is, to be tested in new centers. Moreover, their performance in the hands of less experienced examiners (for whom the models were created) needs to be determined.

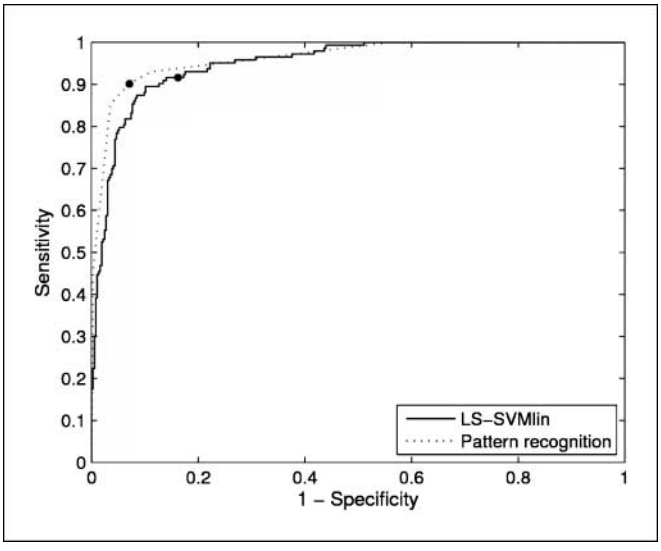

Fig. 1. ROC curve for the best IOTA model LS-SVM lin (11) and pattern recognition (13) based on the level of diagnostic confidence. *Dots,* performance when using the cutoffs (0.15 for LS-SVM lin, prediction of malignancy for pattern recognition).

Table 6. Comparison of performance of the "best" IOTA model and pattern recognition

N = 507	AUC	Cutoff	Sensitivity	Specificity	LR+	LR-
LS-SVM lin (11)	0.950	0.15	91.6%	83.8%	5.65	0.10
Pattern recognition	0.963	Benign or malignant	90.2%	92.9%	12.7	0.11
Difference between LS-SVM lin and pattern recognition	-0.013 (95% CI* = -0.034 to 0.005)		Difference = 1.4% (95% CI, -4.2% to 7.1%)	Difference = -9.1% (95% CI, -13.1% to -5.2%)		

*95% CI was obtained using the bias-corrected bootstrap method (24).

Conclusion

This is the first study to prospectively test the IOTA mathematical models built to predict preoperatively the benign or malignant character of an adnexal mass. Although our results are those of an internal validation, they are reassuring. From all the models that were developed to discriminate between benign and malignant adnexal tumors, the logistic regression models and the LS-SVM with linear and rbf kernel had the highest AUC and were ranked among the five best performing models. However, because the differences in AUCs between the 11 models were extremely small, this might be clinically irrelevant. The models now need to undergo external validation, that is, they need to be tested in completely new centers that did not take part in the IOTA phase 1 study. They should also be tested by ultrasound examiners with varying experience. None of the models tested was superior to subjective evaluation of ultrasound findings by an experienced ultrasound examiner.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References

1. Timmerman D, Schwärzler P, Collins WP, Claerhout F, Coenen M, Amant F. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. Ultrasound Obstet Gynecol 1999;13:11–6.
2. Timmerman D. The use of mathematical models to evaluate pelvic masses; can they beat an expert operator? Best Pract Res Clin Obstet Gynaecol 2004;18:91–104.
3. Valentin L. Prospective cross-validation of Doppler ultrasound examination and gray-scale ultrasound imaging for discrimination of benign and malignant pelvic masses. Ultrasound Obstet Gynecol 1999;14:273–83.
4. Valentin L. Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. Ultrasound Obstet Gynecol 1999;14:338–47.
5. Ferrazzi E, Zanetta G, Dordoni D, Berlanda N, Mezzopane R, Lissoni G. Transvaginal ultrasonographic characterization of ovarian masses: a comparison of five scoring systems in a multicenter study. Ultrasound Obstet Gynecol 1997;10:192–7.
6. Aslam N, Banerjee S, Carr J, Savvas M, Hooper R, Jurkovic D. Prospective evaluation of logistic regression models for the diagnosis of ovarian cancer. Obstet Gynecol 2000;96:75–80.
7. Mol BW, Boll D, De Kanter M, Heintz P, Sijmons E, Oei G. Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. Gynecol Oncol 2001;80:162–7.
8. Van Holsbeke C, Van Calster B, Valentin L, et al. 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;13:4440–7.
9. 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;23:8794–801.
10. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. 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;16:500–5.
11. Van Calster B, Timmerman D, Lu C, et al. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. Ultrasound Obstet Gynecol 2007;29:496–504.
12. Van Calster B, Timmerman D, Nabney I, et al. Using Bayesian Neural Networks with ARD input selection to detect malignant adnexal masses prior to surgery. Neural Comput Appl 2007;17:489–500.
13. Valentin L. Use of morphology to characterize and manage common adnexal masses. Best Pract Res Clin Obstet Gynaecol 2004;18:71–89.
14. Lu C, Suykens JAK, Timmerman D, Vergote I, Van Huffel S. Linear and nonlinear preoperative classification of ovarian tumors. Chapter 11 of knowledge based intelligent system for health care. In: Ichimura T, Yoshida K, editors. Vol. 7. International Series on Advanced Intelligence, Advanced Knowledge International (Magill, Australia); 2004. p. 343–82.
15. Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Machine Learn Res 2001;1:211–44.
16. Pelckmans K, Goethals I, De Brabanter J, Suykens JAK, De Moor B. Componentwise least squares support vector machines. In: Wang L, editor. Support vector machines: theory and applications. Berlin: Springer; 2005. p. 77–98.
17. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. Singapore: World Scientific; 2002.
18. Pochet N, Suykens J. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. Ultrasound Obstet Gynecol 2006;7:607–8.
19. MacKay DJC. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. Netw Comput Neural Syst 1995;6:469–505.
20. Nabney IT. NETLAB. Algorithms for pattern recognition. London: Springer; 2002.
21. Hanley JA, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.
22. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9:190–5.
23. Pepe M, Longton G, Anderson G, Schummer M. Selecting differentially expressed genes from microarray experiments. Biometrics 2003;59:133–42.
24. Chernik MR. Bootstrap methods: a guide for practitioners and researchers. 2nd ed. New York: Wiley; 2007.
25. Newcombe RG. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2. Asymptotic methods and evaluation. Stat Med 2006;25:559–73.
26. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 1998;17:873–90.
27. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med 1998;17:2635–50.
28. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas J. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. BJOG 1990;97:922–9.
29. Royal College of Obstetricians and Gynaecologists guideline no. 34, October 2003.
30. Dearking AC, Aletti GD, McGree M, Weaver A, Sommerfield M-K, Cliby W. How relevant are ACOG guidelines for referral of adnexal mass? Obstet Gynecol 2007;110:841–8.