External Validation of Mathematical Models to Distinguish Between Benign and Malignant Adnexal Tumors: A Multicenter Study by the International Ovarian Tumor Analysis Group

Caroline Van Holsbeke,[1,4] Ben Van Calster,[3] Lil Valentin,[7] Antonia C. Testa,[8] Enrico Ferrazzi,[9] Ioannis Dimou,[5] Chuan Lu,[6] Philippe Moerman,[2] Sabine Van Huffel,[3] Ignace Vergote,[1] and Dirk Timmerman[1]

Abstract

Purpose: Several scoring systems have been developed to distinguish between benign and malignant adnexal tumors. However, few of them have been externally validated in new populations. Our aim was to compare their performance on a prospectively collected large multicenter data set.

Experimental Design: In phase I of the International Ovarian Tumor Analysis multicenter study, patients with a persistent adnexal mass were examined with transvaginal ultrasound and color Doppler imaging. More than 50 end point variables were prospectively recorded for analysis. The outcome measure was the histologic classification of excised tissue as malignant or benign. We used the International Ovarian Tumor Analysis data to test the accuracy of previously published scoring systems. Receiver operating characteristic curves were constructed to compare the performance of the models.

Results: Data from 1,066 patients were included; 800 patients (75%) had benign tumors and 266 patients (25%) had malignant tumors. The morphologic scoring system used by Lerner gave an area under the receiver operating characteristic curve (AUC) of 0.68, whereas the multimodal risk of malignancy index used by Jacobs gave an AUC of 0.88. The corresponding values for logistic regression and artificial neural network models varied between 0.76 and 0.91 and between 0.87 and 0.90, respectively. Advanced kernel-based classifiers gave an AUC of up to 0.92.

Conclusion: The performance of the risk of malignancy index was similar to that of most logistic regression and artificial neural network models. The best result was obtained with a relevance vector machine with radial basis function kernel. Because the models were tested on a large multicenter data set, results are likely to be generally applicable.

There is a need for adequate preoperative assessment of an adnexal mass, because the presumed diagnosis determines the management. Functional cysts and some benign cysts may be treated conservatively. In benign cysts, laparoscopic treatment may decrease hospitalization length and surgical morbidity (1–3). However, if an adnexal mass is likely to be malignant, the patient should be referred to a gynecologic oncologist for surgical staging and optimal debulking. Spilling of an early stage ovarian cancer during laparoscopy or suboptimal debulking with residual disease may worsen the prognosis (4–7).

In most cases, experienced sonologists could correctly determine the character of an adnexal mass on the basis of their own subjective impression. They may achieve a sensitivity with regard to malignancy of >95% and a specificity of ~90% (8, 9). Less experienced sonologists could be helped by scoring systems or mathematical models that can be used to discriminate between benign and malignant adnexal masses.

Authors' Affiliations: Departments of [1]Obstetrics and Gynecology, and [2]Pathology, University Hospitals KU Leuven, [3]Department of Electrical Engineering, ESAT-SCD, Katholieke Universiteit Leuven, Leuven, Belgium; [4]Department of Obstetrics and Gynecology, Ziekenhuis Oost-Limburg, Genk, Belgium; [5]Department of Electronics and Computer Engineering, Technical University of Crete, Chania, Greece; [6]Department of Computer Science, University of Wales, Aberystwyth, United Kingdom; [7]Department of Obstetrics and Gynecology, Malmö University Hospital, Malmö, Sweden; [8]Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del Sacro Cuore, Rome, Italy; and [9]DCS Sacco, Università di Milano, Milan, Italy

Received 12/15/06; revised 3/5/07; accepted 4/2/07.

Grant support: This research was supported by interdisciplinary research grants of the Katholieke Universiteit Leuven, Belgium (IDO/99/03 and IDO/02/09 projects), by the Belgian Programme on Interuniversity Poles of Attraction, by the Concerted Action Project AMBioRICS of the Flemish Community, and by the EU Network of excellence BIOPATTERN into the IST program, entitled "Computational Intelligence for Biopattern Analysis in Support of eHealthcare" (contract no. FP6-2002-IST 508803), and by research grants from the Swedish Medical Research Council (K2001-72X-11605-06A, K2002-72X-11605-07B, K2004-73X-11605-09A, and K2006-73X-11605-11-3), by funds administered by the Malmö General Hospital Foundation for the fight against cancer, and a Swedish governmental grant from the region of Scania.

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.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Requests for reprints: Dirk Timmerman, Department of Obstetrics and Gynecology, University Hospitals, KU Leuven, Herestraat 49, B-3000 Leuven, Belgium. Phone: 32-1634-4201; Fax: 32-1634-4205; E-mail: dirk.timmerman@uz.kuleuven.ac.be.

doi:10.1158/1078-0432.CCR-06-2958

A whole series of prediction models (scoring systems, logistic regression models, and artificial neural networks) have previously been developed (10). Most of them have not been tested prospectively on a large population. Therefore, their generalizibility is unknown.

The aim of this study was to prospectively validate the performance of published scores and mathematical models to predict malignancy in an adnexal mass by applying the models to the data collected in a large multicenter study by the International Ovarian Tumor Analysis (IOTA) group (11).

Materials and Methods

Data set

The data set used was recruited during phase I of the IOTA study. This was the first study from the IOTA group, a prospective multicenter study done in nine centers from five different countries (11). The aim of the IOTA study was to collect sonographic and demographic data of >1,000 patients with an adnexal mass to develop mathematical models to predict malignancy in an ovarian mass. All details have been previously described (11). The primary outcome was the histologic classification of the excised tissue as malignant or benign. Data from 1,066 patients were available for analysis. Serum CA125 values were known for 809 (75.9%) of the 1,066 patients. Models that included CA125 were tested only on these 809 patients. Overall, 800 (75%) tumors were benign and 266 (25%) were malignant.

Selection of scores and mathematical models to predict malignancy

The selection of the old models was partly done by literature search. We selected models for which the statistical formula was published, provided that all the variables used in the formulae were also analyzed in phase I of the IOTA study. In this way, 17 models were selected: three scoring systems [two risk of malignancy indices (RMI and RMI2) and Lerner's scoring system; refs. 12–14], seven logistic regression models (15–21), three artificial neural network models (19, 20), and four least squares support vector machine (LS-SVM) and relevance vector machine (RVM) models (20, 22). LS-SVMs and RVMs are advanced mathematical models that are also highly capable of handling nonlinear data. They are believed to perform well when used prospectively and are briefly described below.

The variables used in the selected scoring systems and models are shown in Table 1. The following models/scores did not fulfill the inclusion criteria: the scoring systems of DePriest et al. (23), Ferrazzi et al. (24), Alcazar et al. (25), and Sassone et al. (26); the logistic regression model of Alcazar et al. (27); and the neural networks from

Table 1. Scores and models tested, variables used in the scores and models, and cutoffs of the respective scores and models suggested in the original report

Reference	Type of model	Variables used	Cutoff*
RMI (12)	Scoring system	Menopausal score, CA125, multilocular cysts, solid areas, metastases, ascites, bilaterality	200
RMI2 (13)	Scoring system	Menopausal score, CA125, multilocular cysts, solid areas, metastases, ascites, bilaterality	4
Lerner score (14)	Scoring system	Wall structure, acoustic shadows, septa, echogenicity	4
Minaretzis et al. (15)	LR	Tumor consistency (multilocular or solid = 1), bilaterality, diameter, age	—
Tailor et al. (16)	LR	Papillations, age, time-averaged maximum mean velocity	50%
Prömpeler et al. (17)	LR	Ascites, solid lesion without shadowing, cyst with >30% solid part, diameter of the lesion, multilocularity, surface of the cyst	10%
Timmerman et al. (18)	LR	Color score, CA125, papillations, menopausal score	25%
Timmerman et al. (19)	LR	Papillations, internal wall, unilocular cyst, ascites, bilaterality, menopausal score, CA125	60%
Lu et al. (20)	LR	CA125, papillations, solid tumor, color score of at least 3, bilaterality, menopausal status, ascites, acoustic shadows, color score 4, irregular cyst wall	20%
Jokubkiene et al. (21)	LR	Size of lesion (mean of three diameters), size of solid part (mean of three diameters), irregular wall	12%
Timmerman et al., NN1 (19)	ANN	Papillations, color score, menopause, CA125	45%
Timmerman et al., NN2 (19)	ANN	Papillations, smooth surface, unilocularity, ascites, bilaterality, menopause, CA125	60%
Bayesian MLP, Lu et al. (20)	ANN	CA125, papillations, solid tumor, color score of at least 3, bilaterality, menopausal status, ascites, acoustic shadows, color score 4, irregular wall	30%
SVM linear kernel, Lu et al. (22)	Bayesian LS-SVM	CA125, papillations, solid tumor, color score of at least 3, bilaterality, menopausal status, ascites, acoustic shadows, color score 4, irregular wall	30%
SVM RBF kernel, Lu et al. (22)	Bayesian LS-SVM	CA125, papillations, solid tumor, color score of at least 3, bilaterality, menopausal status, ascites, acoustic shadows, color score 4, irregular wall	30%
RVM linear kernel, Lu et al. (22)	RVM	CA125, papillations, solid tumor, color score of at least 3, bilaterality, menopausal status, ascites, acoustic shadows, color score 4, irregular wall	30%
RVM RBF kernel, Lu et al. (22)	RVM	CA125, papillations, solid tumor, color score of at least 3, bilaterality, menopausal status, ascites, acoustic shadows, color score 4, irregular wall	30%

Abbreviations: MLP, multilayer perceptron; ANN, artificial neural network; LR, logistic regression.
*The cutoff values refer to the output value of the scoring system or mathematical model at or above which the tumor is classified as malignant.

Biagiotti et al. (28), Clayton et al. (29), Bishop (30), and Tailor et al. (31). The reasons for exclusion are listed in the Supplementary Table that is published online.

LS-SVMs. SVMs are learning machines that, using a kernel function, nonlinearly transform the input space into a high dimensional feature space. In this high dimensional feature space, a linear classifier is constructed. Depending on the type of transformation (linear or nonlinear kernel), this linear classifier in the feature space coincides with either a linear or nonlinear classifier in the original input space. The SVM looks for a linear classifier in the feature space by maximizing the margin between the data of both classes (leading to simpler models). Therefore, SVMs look for a trade-off between minimizing complexity and minimizing the number of misclassifications. The resulting model is sparse in that only a few cases from the training set were used to construct the decision boundary. These cases are called the "support vectors" and are usually located on or close to the decision boundary. The LS-SVM is a variant of the standard SVM in which training of the model is greatly simplified (20, 32). Lu et al. developed a linear (using a linear kernel) and a nonlinear (using a radial basis function or RBF kernel to transform the input data) LS-SVM model in a Bayesian framework (20, 22).

RVMs. RVMs were inspired by the SVM model formulation, but because they do not work with a feature space, there are less restrictions with respect to valid transformations of the input space (32). Therefore, they are still fundamentally different from SVMs. RVM models use a Bayesian perspective and the resulting model is again sparse (33). However, the cases used for constructing the decision boundary were prototypical examples of both classes rather than cases that were close to the decision boundary. Lu et al. developed a linear and a nonlinear (using an RBF kernel) RVM model (20, 22). The principles of logistic regression analysis, artificial neural networks, SVMs, and RVMs are described in refs. (20, 30, 32–37).

Modification of variables

Some variables used in some scores or models lacked an exact corresponding variable in the IOTA database. The IOTA variables that we used in these cases are shown in the Supplementary Tables published online.

Statistical analysis

Statistical analyses were carried out using Statistical Analysis Software version 9.1 (SAS Institute, Inc.). The performance of the models and scores were compared with respect to their area under the receiver operating characteristic (ROC) curve (AUC). After applying the cutoff value to predict malignancy suggested in the original publication, the sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio (LR+), and negative likelihood ratio (LR-) for each score/model could be calculated. We used the cutoffs reported in the original articles, although the optimal cutoff was highly dependent on the population characteristics, e.g., the amount of malignant and exceptional tumors. Because the sensitivity and related measures were dependent on the cutoff value chosen, whereas the AUC was not, we considered the AUC to be the most important measure of diagnostic performance (38, 39). The nonparametric procedure described in ref. (38) was used to test whether the AUCs of two models differed. To compare all 17 scores/models with each other, we used a ranking method based on the work of Pepe et al. (40). All models were ranked with respect to a chosen criterion. This ranking, however, is influenced by sampling variability. This variability is accounted for by computing the probability that a method is ranked among the κ best

Table 2. Diagnostic performance of scores and models tested

Method	AUC in original report on test set	Prospective testing on IOTA data set, all cases (n = 1,066)		Prospective testing on IOTA data set, CA125 cases (n = 809)	
		AUC	SE (95% CI)	AUC	SE (95% CI)
Scoring systems					
RMI (12)	—			0.883	0.013 (0.858-0.908)
RMI2 (13)	—			0.860	0.016 (0.829-0.891)
Lerner et al. (14)	—	0.664	0.017 (0.631-0.697)	0.676	0.019 (0.639-0.713)
Logistic regression					
Minaretzis et al. (15)	—	0.827	0.013 (0.802-0.852)	0.813	0.015 (0.784-0.842)
Tailor et al. (16)	1.000	0.857	0.013 (0.832-0.882)	0.856	0.014 (0.829-0.883)
Prömpeler et al. (17)					
All women	—	0.795	0.015 (0.766-0.824)	0.788	0.017 (0.742-0.821)
Premenopausal*	—	0.799	0.026 (0.748-0.850)	0.800	0.028 (0.745-0.855)
Postmenopausal*	—	0.724	0.024 (0.677-0.771)	0.720	0.026 (0.669-0.771)
Timmerman et al. (18)	0.904			0.890	0.012 (0.866-0.914)
Timmerman et al. (19)	0.965			0.894	0.012 (0.870-0.918)
Lu et al. (20)	0.911			0.910	0.011 (0.888-0.932)
Jokubkiene et al. (21)	0.98 †	0.894	0.011 (0.872-0.916)	0.895	0.012 (0.871-0.919)
Artificial neural networks					
NN1, Timmerman et al. (19)	0.951			0.889	0.013 (0.864-0.914)
NN2, Timmerman et al. (19)	0.983			0.876	0.014 (0.849-0.903)
Bayesian MLP, Lu et al. (20)	0.917			0.890	0.013 (0.865-0.915)
Bayesian LS-SVMs					
Linear kernel, Lu et al. (22)	0.914			0.916	0.011 (0.894-0.938)
RBF kernel, Lu et al. (22)	0.918			0.911	0.011 (0.899-0.933)
RVMs					
Linear kernel, Lu et al. (22)	0.92			0.921	0.011 (0.899-0.943)
RBF kernel, Lu et al. (22)	0.92			0.923	0.011 (0.901-0.945)

NOTE: CA125 refers to the population of 809 women with CA125 values available; All refers to the total population of 1,066 patients.
Abbreviations: 95% CI, 95% confidence intervals; NN, neural network; MLP, multilayer perceptron.
*Six hundred and thirty-five patients were premenopausal, CA125 values were available for 445 patients; 431 patients were postmenopausal, CA125 values were available for 364 patients.
† In the original report, there was no split of the population into training and test sets, therefore, the AUC for the total population is reported.

methods: Pm(κ) (ref. 40). One thousand bootstrap samples were drawn using the IOTA phase I data, which have CA125 to allow the comparison of all models. For each bootstrap sample, the criterion was computed for each of the 17 models and the models were ranked within each bootstrap from good to bad. This allowed us to compute how many times each model was ranked among the best 5 and best 10 in order to obtain Pm(5) and Pm(10). As already mentioned, the main criterion used was the AUC. However, based on the work of Pepe et al., we also looked at the partial AUC (pAUC; ref. 40). The pAUC or the partial AUC gives you the area under a part of the ROC curve with a minimum acceptable specificity. We believe that the minimum acceptable specificity level when predicting the malignancy of ovarian tumors is 75%; therefore, we computed the partial AUC (0.75) for that part of the ROC curve in which the specificity is at least 75% (i.e., the most left part).

Results

The diagnostic performance of the scores and models tested is shown in Tables 2 and 3 and in Fig. 1. The scoring system that did best when applied on the IOTA data was the RMI of Jacobs et al. (12) with an AUC of 0.88. The AUCs of the scores ranged from 0.66 to 0.88 (Tables 2 and 3; Fig. 1A). The AUCs of the logistic regression models ranged from 0.72 to 0.91 with the logistic regression model of Lu et al. (20) performing best (AUC, 0.91; Tables 2 and 3; Fig. 1B). All neural networks did well, their AUCs ranging from 0.88 to 0.89, and so did the LS-SVMs and RVMs with AUCs of 0.91 to 0.92 (Tables 2 and 3; Fig. 1C).

Figure 1D shows the best performing model per category, the RVM with RBF kernel of Lu et al. (22) having the highest AUC (0.923). Table 4 shows the 17 scores and models ranked from high to low AUC. The five best performing models were the RVMs, the LS-SVMs, and the logistic regression model from Lu et al. (20, 22). By using the ranking method from Pepe et al. (40), the five best performing models were the same (Table 4). The probability of being among the best five models, Pm(5), or the best 10 models, Pm(10), is much higher for these models than for any other model. The other criterion used, the pAUC(0.75), also ranked these five models at the top (Table 4).

Discussion

This is the first report on prospective testing of multiple mathematical models to predict malignancy in an adnexal mass from a large and multicenter database (41–43). External validation (i.e., prospective testing of a model on a new database in another center) is the ultimate test a model has to go through before it can be used in clinical practice. Internal validation or testing the performance of the model on a test set of data from the same center is less adequate because it does not give information on the generalizibility of the model. The advantage of the IOTA database is that data were collected in different

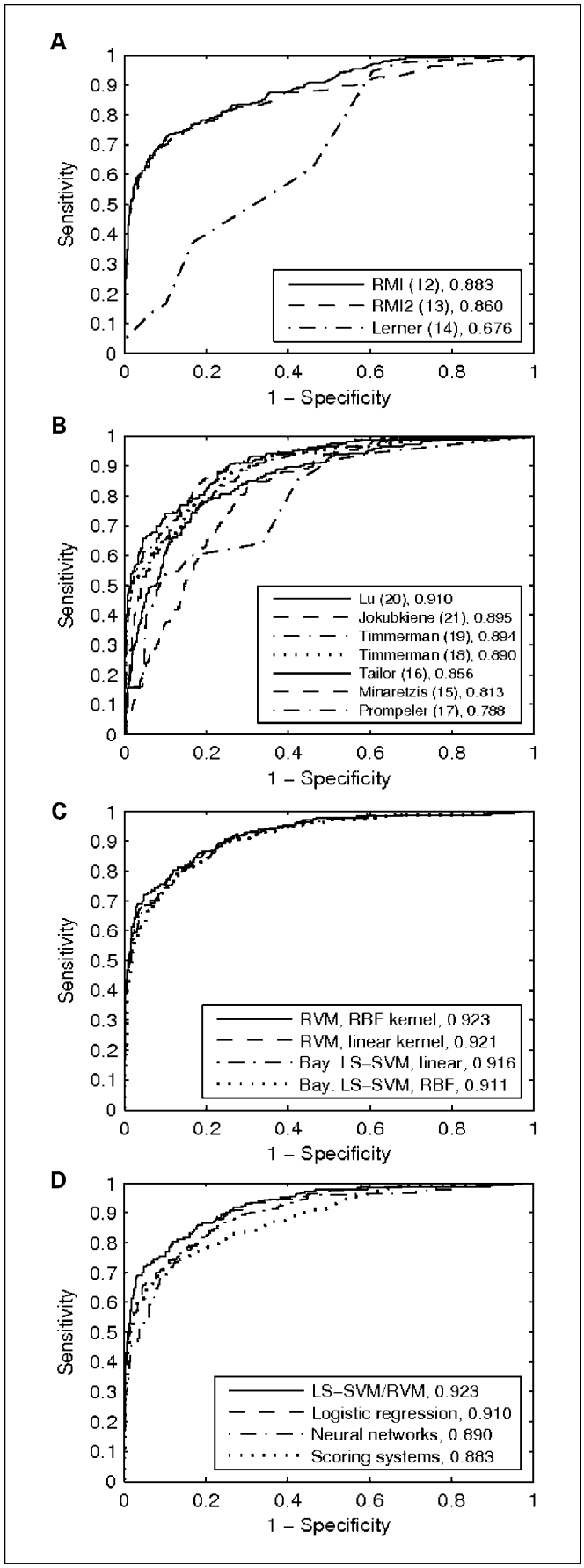

Fig. 1. *A,* AUC of two RMIs and of the Lerner score when tested on the 809 patients with available CA125 results. *B,* AUC of seven logistic regression models when tested on the 809 patients with available CA125 results. *C,* AUC of two Bayesian LS-SVM and two RVM when tested on the 809 patients with available CA125 results. *D,* AUC of the best diagnostic model per category when tested on the 809 patients with available CA125 results. For the scoring systems, the best score is the RMI (12); for the neural networks, this is the Bayesian MLP (multi-layer perceptron) from Lu et al. (20); for the logistic regression models, the model from Lu et al. (20) and for the vector machine models, the RVM with RBF kernel from Lu et al. (20, 22).

Table 3. Sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios for all models when using the cutoff value suggested in the original report in comparison with the originally reported sensitivity and specificity for the test set patients

Method	Original report				Cutoff
	Total population	Test set patients	Sensitivity test set	Specificity test set	
Scoring systems					
RMI (12)	143	—	85*	97*	200
RMI2 (13)	173	—	80*	92*	200
Lerner et al. (14)	350	—	96.8*	77*	4
Logistic regression					
Minaretzis et al. (15)	959	—	—	—	—
Tailor et al. (16)	52	15	100	100	25%
	52	15	100	100	50%
Prömpeler et al. (17)	754	—	—	—	10%
Prömpeler et al. (17), premenopausal	400	—	86.5*	92.6*	10%
Prömpeler et al. (17), postmenopausal	354	—	93*	82.7*	10%
Timmerman et al. (18)	173	57	92.9	78.4	25%
Timmerman et al. (19)	173	57	93.8	83	60%
Lu et al. (20)	265	160	81.5	80.2	50%
Jokubkiene et al. (21)	106	—	100*	90*	12%
Artificial neural networks					
NN1, Timmerman et al. (19)	173	57	87.5	92.7	45%
NN2, Timmerman et al. (19)	173	57	93.8	95.1	60%
Bayesian MLP, Lu et al. (20)	265	160	83.3	81.1	30%
Bayesian LS-SVMs					
Linear kernel, Lu et al. (22)	265	160	83.3	81.1	30%
RBF kernel, Lu et al. (22)	265	160	85.2	84	30%
RVMs					
Linear kernel, Lu et al. (22)	265	160	85.2	82.1	30%
RBF kernel, Lu et al. (22)	265	160	87	82.1	30%

Abbreviations: LR+, positive likelihood ratio; LR-, negative likelihood ratio; NN: neural network; MLP: multilayer perceptron.
*In the original report, there was no split of the population into a training and test set, therefore, sensitivity and specificity for the total population is reported. CA125 refers to the population of 809 women with CA-125 values available; All refers to the total population of 1066 patients.
† Six hundred and thirty-five patients were premenopausal, CA125 values were available for 445 patients; 431 patients were postmenopausal, CA125 values were available for 364 patients.

centers with different referral patterns resulting in a diverse study population mimicking a more universal population.

Most models tested on the IOTA database did worse than was originally reported. For an excellent model, the AUC should reach 0.90 (35). Only the logistic regression model of Lu and colleagues (20), the LS-SVM with linear and RBF kernel, and the RVM with linear and RBF did so (20, 22).

There are several plausible reasons why the performance in the original articles was better than the results that we report. One is that some variables in some scores and models lacked an exact corresponding variable in the IOTA database, so that the variables had to be redefined on the basis of the information available in the IOTA database. This was true for both RMIs, the Lerner score, and the Prömpeler logistic regression model. Second, one of the causes of poorer performance in a prospective test are the differences in population characteristics between the data from the original study and the data from the prospective study. If the model development was based on a small sample of tumors, the particularities of that specific sample will have great influence on the model. However, Table 3 shows the population size in the original reports and we could not find a linear correlation between the size of the data set in the original report and the performance of the score or the model. The prevalence of malignancy in a population and the mix of exceptional tumors and borderline tumors will characterize a population. The prevalence of malignancy, for example, is 25% in the IOTA phase I study, but when we look at the original reports of the models that were tested, it varied between 22% and 42%. Models that have been developed on a large and multicenter study sample, representative of a general population, are likely to be more robust. Therefore, a model developed in a tertiary referral center might perform less when it is applied in a regional gynecology center. Because the IOTA data set is a multicenter data set collected in hospitals from different countries and with different referral pattern, it is more likely to represent a "universal" population.

A third explanation for models incorporating ultrasound variables performing worse when tested prospectively, is that

Table 3. Sensitivity, specificity, positive and negative predictive values, and positive and negative likelihood ratios for all models when using the cutoff value suggested in the original report in comparison with the originally reported sensitivity and specificity for the test set patients (Cont'd)

Prospective testing on IOTA data						
Study population	**Sensitivity**	**Specificity**	**Positive predictive value**	**Negative predictive value**	**LR+**	**LR-**
CA125	69	91.7	78.0	87.4	8.31	0.34
CA125	75.2	84.8	67.9	88.9	4.95	0.29
CA125	62	54.1	36.6	76.9	1.35	0.70
All	62	51.4	29.8	80.3	1.28	0.74
—	—	—	—	—	—	—
CA125	60.7	90.1	72.4	84.3	6.13	0.44
All	59.8	90	66.5	87.1	5.98	0.45
CA125	49.2	93.5	76.3	81.2	7.57	0.54
All	48.5	93.8	72.1	84.6	7.82	0.55
CA125	59.9	84.0	61.4	83.1	3.74	0.48
All	57.5	86.0	57.7	85.9	4.11	0.49
CA125 †	57.0	92.8	65.3	90.0	7.92	0.46
All †	53.6	93.7	60.5	91.8	8.51	0.50
CA125 †	61.5	68.8	59.6	70.4	1.97	0.56
All †	59.8	70.3	56.4	73.1	2.01	0.57
CA125	78.1	79.2	61.6	89.4	3.75	0.28
CA125	83.9	75.3	59.2	91.6	3.40	0.21
CA125	89.3	75.0	60.3	94.2	3.57	0.14
CA125	81.4	83.4	67.7	91.3	4.90	0.22
All	78.9	84.6	63.1	92.4	5.12	0.25
CA125	81.4	77.8	61.0	90.7	3.67	0.24
CA125	98.4	34.0	38.9	98.0	1.49	0.05
CA125	67.8	90.8	75.9	86.9	7.37	0.35
CA125	90.1	75.1	60.7	94.7	3.62	0.13
CA125	89.3	76.2	61.5	94.3	3.75	0.14
CA125	90.5	74.6	60.3	94.8	3.56	0.13
CA125	90.1	74.3	59.9	94.6	3.51	0.13

the results of some ultrasound variables may be difficult to reproduce, in particular, those including an element of subjectivity, e.g., regular versus irregular cyst wall and those highly dependent on the examination technique adopted, e.g., Doppler variables. Results of spectral Doppler ultrasound examinations depend on the tumor vessel investigated, velocity variables like peak systolic velocity and time-averaged maximum velocity are angle-dependent, and estimation of the color content of the tumor scan to assess tumor vascularity is subjective. This may make scores and models incorporating Doppler variables difficult to reproduce.

In the IOTA database, all variables had been defined using strict criteria, but these criteria may not have corresponded exactly to those used when creating the score or model to be tested. The models that did best in this study, all had been developed on data that had—at least partly—been collected by examiners who also collected data for the IOTA study, and all had been created at institutions that contributed data to the IOTA study. This means that the variables used in these models were probably defined similarly when the models were created and tested, and that the tumor populations in which the models were created and tested were also similar. This may partly explain the apparent superior performance of these models. When using the AUC as a measure of diagnostic performance, the performances of the RMI of Jacobs et al. (12) and the artificial neural network models from Timmerman et al. (19) were surprisingly similar (AUC, 0.883-0.876-0.889), even with the very best mathematical model, the maximum difference in AUC was only 0.04 (0.883 versus 0.923). Also, in other studies, the RMI of Jacobs has been shown to perform very well when tested prospectively (13, 18, 44, 45). The RMI was developed on a relatively small population (143 patients) but its robustness may be explained by the inclusion of variables that did not require high ultrasound skills except for the diagnosis of metastatic lesions (CA125, menopausal status, tumor with any other ultrasound morphology than a unilocular cyst, ascites, bilateral lesions, and presence of metastases at a scan). On the other hand, it requires the knowledge of the serum CA125 level, which is expensive and time-consuming and not always available at the time the scan is done. The RMI is one of the oldest multimodal scoring systems that uses the combination of a tumor marker and ultrasound variables.

When we used the cutoffs that were reported in the original articles, the RBF kernel model picked up another 51 additional malignancies in comparison with the RMI (218 of 242 versus 167 of 242 malignant masses) due to higher sensitivity (Table 4). Furthermore, when using the ranking method of Pepe et al. (40), the RMI was not listed in the top five best performing models.

Finally, the statistical development procedure of the models is of utmost importance. The final model depends on the

Table 4. All 17 scoring systems and mathematical models ranked from highest to lowest AUC

Model	Type of model	AUC	Pm(5)	Pm(10)	pAUC(0.25)
RVM (RBF kernel), Lu et al. (22)	RVM	0.923	1.000	1.000	0.193
RVM (linear kernel), Lu et al. (22)	RVM	0.921	1.000	1.000	0.191
Bay LS-SVM (linear kernel), Lu et al. (22)	Bayesian LS-SVM	0.916	1.000	1.000	0.186
Bay LS-SVM (RBF kernel), Lu et al. (22)	Bayesian LS-SVM	0.911	0.954	1.000	0.183
Lu et al. (20)	LR	0.910	0.912	1.000	0.181
Jokubkiene et al. (21)	LR	0.895	0.095	0.878	0.167
Timmerman et al. (18)	LR	0.894	0.021	0.985	0.171
Timmerman et al. (19)	LR	0.890	0.000	0.904	0.168
Bay MLP, Lu et al. (20)	ANN	0.890	0.006	0.790	0.172
NN1, Timmerman et al. (19)	ANN	0.889	0.003	0.864	0.169
RMI (12)	Scoring system	0.883	0.009	0.475	0.174
NN2, Timmerman et al. (19)	ANN	0.876	0.000	0.086	0.165
RMI2 (13)	Scoring system	0.860	0.000	0.003	0.172
Tailor et al. (16)	LR	0.856	0.000	0.015	0.150
Minaretzis et al. (15)	LR	0.813	0.000	0.000	0.105
Prömpeler et al. (17)	LR	0.788	0.000	0.000	0.119
Lerner score (14)	Scoring system	0.676	0.000	0.000	0.063

NOTE: Pm(5) gives the chance that the model is one of the five best performing models; Pm(10) gives the chance that the model is one of the 10 best performing models; pAUC(0.75) gives the partial AUC under the ROC curve in the area with a specificity of at least 75%; in this respect, the maximum pAUC(0.75) can be only 0.25.
Abbreviations: LR, logistic regression model; MLP, multilayer perceptron; ANN: artificial neural network.

selection of the different variables. The goodness of fit is important because it implies the accuracy with which the final regression model describes the data (46). Although the AUC represents the quality of a test and is very useful to compare the performance of different models, the selection of cutoff levels is dependent on the study population and the preference of the investigator: if a high sensitivity is chosen, few malignancies will be missed, whereas increasing the cutoff will lead to a higher specificity and thus fewer false-positive results. It is interesting to note that the sensitivities and specificities of the scores/models when using the cutoffs that had been recommended in the original publications were not very good, and they were much worse than in the original publications. It is well known that the optimal cutoff point of any test will depend on the proportion of cancers in the study population; therefore, if one uses the same cutoff in a population with a significantly different amount of malignancies, one can expect that the model will perform worse (Table 3).

We conclude that the AUC of most logistic regression and artificial neural network models was similar. The best results were obtained with vector machine models. The use of these complex mathematical models resulted in the correct classification of a significant amount of additional malignancies. Because the models were tested on a large multicenter data set, results are likely to be generally applicable.

Appendix A

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, Naples, Italy
Antonia Testa, Rome, 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, Monza, Italy.tr

References

1. Buchweitz O, Matthias S, Muller-Steinhardt M, Malik E. Laparoscopy in patients over 60 years old: a prospective randomized evaluation of laparoscopic versus open adnexectomy. Am J Obstet Gynecol 2005; 193:1364–8.
2. Medeiros LR, Fachel JM, Garry R, Stein AT, Furness S. Laparoscopy versus laparotomy for benign ovarian tumours. Cochrane Database Syst Rev 2005;20: CD004751.
3. Carley ME, Klingele CJ, Gebhart JB, Webb MJ, Wilson TO. Laparoscopy versus laparotomy in the management of benign unilateral adnexall masses. J Am Assoc Gynecol Laparosc 2002;9:321–6.
4. Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P. Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. Lancet 2001;357:176–82.
5. Hacker NF, Berek JS, Lagasse LD. Primary cytoreductive surgery for epithelial ovarian cancer. Obstet Gynecol 1983;61:431–20.
6. Sutton GP, Stehman FB, Einhorn LH. Ten-year follow-up of patients receiving cisplatin, doxorubicin, and cyclophosphamide chemotherapy for advanced epithelial ovarian carcinoma. J Clin Oncol 1989;7:223–9.
7. Redman JR, Petroni GR, Saigo PE. Prognostic factors in advanced ovarian carcinoma. J Clin Oncol 1986;4: 515–23.
8. 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.
9. 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.

10. 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.

11. 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.

12. 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.

13. Tingulstad S, Hagen B, Skjeldestad FE, Onsrud M, Kiserud T, Halvorsen T. Evaluation of a risk of malignancy index based on serum CA 125, ultrasound findings and menopausal status in the preoperative diagnosis of pelvic masses. BJOG 1996; 103:826–31.

14. 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;170:81–5.

15. Minaretzis D, Tsionou C, Tziortziotis D, Michalas S, Aravantinos D. Ovarian tumors: prediction of the probability of malignancy by using patient's age and tumor morphologic features with a logistic model. Gynecol Obstet Invest 1994;38:140–4.

16. Tailor A, Jurkovic D, Bourne T, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis. Ultrasound Obstet Gynecol 1997;10:41–7.

17. Prömpeler HJ, Madjar H, Sauerbrei W, Lattermann U, Pfleiderer A. Diagnostic formula for the differentiation of adnexal tumors by transvaginal sonography. Obstet Gynecol 1997;89:428–33.

18. Timmerman D, Bourne T, Tailor A, Collins WP, Verrelst H, Vandenberghe K. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model. Am J Obstet Gynecol 1999;181:57–65.

19. Timmerman D, Verrelst H, Bourne TH, De Moor B, Collins WP, Vergote I. Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses. Ultrasound Obstet Gynecol 1999;13:17–25.

20. Lu C, Suykens JAK, Timmerman D, Vergote I, Van Huffel S. Linear and nonlinear preoperative classification of ovarian tumors. Chapter 11, Knowledge based intelligent system for health care. In: Ichimura T, Yoshida K, editors. Vol. 7, International Series on Advanced Intelligence. Magill, Australia: Advanced Knowledge International; 2004. p. 343–82.

21. Jokubkiene L, Sladkevicius P, Valentin L. Does three-dimensional power Doppler ultrasound help in discrimination between benign and malignant ovarian masses? Ultrasound Obstet Gynecol 2007;29: 215–25.

22. Lu C, Van Gestel T, Suykens JAK, Van Huffel S, Vergote I, Timmerman D. Preoperative prediction of malignancy of ovarium tumor using least squares support vector machines. Artif Intell Med 2003;28: 281–306.

23. DePriest PD, Varner E, Powell J, Fried A, Puls L, Higgins R. The efficacy of a sonographic morphology index in identifying ovarian cancer: a multi-institutional investigation. Gynecol Oncol 1994;55:174–8.

24. 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.

25. Alcazar JL, Mercé L, Laparte C, Jurado M, Lopez-Garcia G. A new scoring system to differentiate benign from malignant adnexal masses. Am J Obstet Gynecol 2003;188:685–92.

26. Sassone AM, Timor-Tritsch I, Artner A, Westhoff C, Warren W. Transvaginal sonographic characterization of ovarian disease: evaluation of a new scoring system to predict ovarian malignancy. Obstet Gynecol 1991; 78:70–6.

27. Alcazar JL, Jurado M. Using a logistic model to predict malignancy of adnexal masses based on menopausal status, ultrasound morphology and color Doppler findings. Gynecol Oncol 1998;69:146–50.

28. Biagiotti R, Desii C, Vanzi E, Gacci G. Predicting ovarian malignancy: application of artificial neural networks to transvaginal and color Doppler flow US. Radiology 1999;210:399–403.

29. Clayton RD, Snowden S, Weston MJ, Mogensen O, Eastaugh J, Lane G. Neural networks in the diagnosis of malignant ovarian tumours. BJOG 1999;106: 1078–82.

30. Bishop CM. Neural networks for pattern recognition. Oxford University Press; 1996.

31. Tailor A, Jurkovic D, Bourne TH, Collins WP, Campbell S. Sonographic prediction of malignancy in adnexal masses using an artificial neural network. BJOG 1999;106:21–30.

32. Suykens J, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least square support vector machines. Singapore: World Scientific.

33. MacKay D. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. Network: Computation in Neural Systems 1995;6:469–505.

34. 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.

35. Hosmer DW, Lemeshow S. Applied logistic regression. Wiley Series in Probability and Statistics. New York; 2000.

36. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.

37. Tipping ME. Sparse Bayesian learning and the relevance vector machine. J Mach Learn Res 2001;1: 211–44.

38. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves. A nonparametric approach. Biometrics 1988;44:837–45.

39. Hanley JA, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

40. Pepe M, Longton G, Anderson G, Schummer M. Selecting differentially expressed genes from microarray experiments. Biometrics 2003;59:133–42.

41. 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.

42. Valentin L. Comparison of Lerner score, Doppler ultrasound examination, and their combination for discrimination between benign and adnexal masses. Ultrasound Obstet Gynecol 2000;15: 143–7.

43. 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: 357–65.

44. Davies AP, Jacobs I, Wolas R, et al. The adnexal mass: benign or malignant? Evaluation of a risk of malignancy index. BJOG 1993;100:927–31.

45. Morgante G, la Marca A, Ditto A, De Leo V. Comparison of two malignancy risk indices based on serum CA125, ultrasound score and menopausal status in the diagnosis of ovarian masses. BJOG 1999;106: 524–7.

46. Khan K, Chien P, Dwarakanath L. Logistic regression models in obstetrics and gynecology literature. Obstet Gynecol 1999;93:1014–20.