PROSPECTIVE INTERNAL VALIDATION OF MATHEMATICAL MODELS TO PREDICT MALIGNANCY IN ADNEXAL MASSES: RESULTS FROM THE INTERNATIONAL OVARIAN TUMOR ANALYSIS (IOTA) STUDY

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

1 Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium
2 Department of Obstetrics and Gynecology, Ziekenhuis Oost-Limburg, Genk, Belgium
3 Department of Electrical Engineering, K.U.Leuven, Leuven, 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
6 Department of Obstetrics and Gynecology, Malmö University Hospital, Lund University, Malmö, Sweden

Corresponding author: Dirk Timmerman, MD, PhD
Department of Obstetrics and Gynecology,
University Hospitals, KU Leuven,
Herestraat 49, B-3000 Leuven, BELGIUM.
Telephone: + 32 16 344201 (office) Fax: + 32 16 344205 (office)
Telephone: + 32 16 344216 (secretary)
E-mail: dirk.timmerman@uz.kuleuven.ac.be

Keywords: ultrasonography, ovarian neoplasms, color Doppler sonography, logistic models

Running Head: Prospective evaluation of mathematical models

Statement of clinical relevance

For clinicians, ultrasound pattern recognition is the standard method for the preoperative prediction of malignancy in an adnexal mass. IIn 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 (ref Timmerman interobeserver repeoducibility). 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 (ref!) –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 International Ovarian Tumor Analysis study, the IOTA study) including 1066 patients scanned using same standardized ultrasound protocol (ref). In phase 1 of the IOTA study all models proved to perform very well with area under the receiver operating characteristic curve (AUC) of more than 0.92. This area is larger than that of the Risk of Malignancy Index, RMI (ref), that is often regarded as a standard test (9, 11, 12, 30, 31). Internal validation is the next step in the evaluation process of new diagnostis tests. This manuscript describes the internal validation of our new modles. 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 valiadtion of the models in the hands of less experienced examiners remains to be done.

Abstract

Aim

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 dataset on a new dataset, 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 histological classification of excised tissue. The diagnostic performance of 11 mathematical models that had been developed on the phase 1 dataset and of pattern recognition was expressed as AUC and as sensitivity and specificity when using the cut-offs recommended in the studies where the models had been created. For pattern recognition an AUC was made on the basis of 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 % and AUCs between 0.945 and 0.950.

A least square support vector machine with linear kernel and a logistic regression model had the largest AUCs. For pattern recognition, the AUC was 0.963, sensitivity 90.2% and specificity 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 dataset. Pattern recognition used by an expert sonologist remains the best method although the difference in performance between the best mathematical model is not large.

Introduction

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 demonstrate 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, i.e., 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 dataset collected by the International Ovarian Tumor Analysis (IOTA) collaborative group, the IOTA phase 1 dataset (8). Unfortunately, most models and scoring systems performed worse than they had done in the studies 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 in order to create new scoring systems and

mathematical models to distinguish benign from malignant adnexal tumors (9). More than 50 variables of 1066 patients with 1233 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 dataset, and to compare their performance with that of pattern recognition which is assumed to be the standard method.

Methods

New prospective dataset

Patients were recruited from three ultrasound centers: University Hospitals Leuven, Leuven, Belgium; Malmö University Hospital, Malmö, Sweden, and Instituto di Clinica Ostetrica e Ginecologica, Rome, Italy. These centers had contributed two thirds of the cases to the IOTA phase 1 dataset (25%, 30% and 12%, respectively). Consecutive patients meeting the inclusion criteria of the IOTA phase I 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: 1° classification of each mass as benign or malignant on the basis of subjective evaluation of ultrasound findings (pattern recognition), 2° 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 (13) were used for pattern recognition.

Pattern recognition and the following mathematical models developed on the IOTA phase 1 dataset were tested on the newly collected study population: two logistic regression models (LR1 (12 variables) and LR2 (6 variables)) (9), three support vector machine

models (LS-SVM lin, LS-SVM rbf, LS-SVM add rbf (11), three relevance vector machine models (RVM lin, RVM rbf, RVM add rbf (11) and three neural networks (Bay MLP 11-2a, Bay MLP 11-2b and Bay Perc 11 (12).

LR1 was developed using statistical state-of-the-art logistic regression modeling, with due attention to multicollinearity, possible interactions, and the 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 area under the ROC curve (AUC) of 0.942 on the test set (n = 312). Sensitivity and specificity were 93% and 76% using a cut-off of 0.10. LR1 was the main logistic regression model. A reduced model (LR2) using only six variables was developed as well (9).

Kernel methods were also applied for binary classification purposes in (12, 14-16). Methods considered are Bayesian least squares support vector machines (LS-SVMs) (17) and relevance vector machines (RVMs) (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 non-linear 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 while 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 (11, 17, 18). More information about the Bayesian approach to LS-SVMs is given in (11). Variable selection was performed 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 ranging from 83 to 86%, sensitivity from 91 to 93% and specificity from 81 to 84%.

Next to the kernel methods, a Bayesian multi-layer perceptron (MLP) using the evidence procedure (19, 20) was applied as another type of model that can incorporate non-linearity (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 MLP included 11 variables with 2 hidden neurons (Bay MLP 11-2a). All three neural networks used different sets of variables; the models included six to twelve of the variables which are listed in table 1.

Statistical analysis

The diagnostic performance of each model was expressed as the area under the ROC curve, AUC (21) and as the partial AUC (pAUC) (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), i.e., the area under that part of the ROC curve where the specificity is 75% or more (the most left part of the ROC curve) (21). Using the six levels of diagnostic confidence as different cut-offs, an 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 (PPV and NPV), and positive and negative likelihood ratio (LR+ and LR-) when using the risk cut-off recommended in the original article describing the model. However, sensitivity and specificity of a model depends on the cut-off 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, $Pm(\kappa)$, is computed using 1000 bootstrap samples (23). A bootstrap sample is a new dataset generated from the original dataset, where both have the same sample size. The bootstrap sample is constructed by randomly selecting patients from the original dataset. Patients are selected 'with replacement', meaning that a patient can be selected more than once or not at all for a given bootstrap sample (24). 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) was used as the main indicator (25). 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 malignant group. 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 receiver operating characteristic curve analysis has different objectives (25). Nonetheless, this relationship may help non-statisticians 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% confidence interval 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% (143/507), with a malignancy rate of 30% (85/287) in Leuven, 33% (32/96) in Rome, and 21% (26/124) in Malmö. Histological diagnoses are shown in Table 2. Table 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 ranging from 91.6% to 95.1 %, and specificity from 73.9% to 83.8 %. A least square support vector machine 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 pAUC(0.75), i.e., 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 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 92.9%. This corresponds to a positive LR of 12.7 and a negative LR of 0.11 (Table 5 and figure 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% (39/507) of the cases and completely confident about the benign or malignant character of the mass in 46% (226/507) 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 figure 1 compare the performance of the best IOTA model (LS-SVM lin (11) with the assumed standard method, pattern recognition. The AUC for LS-SVM lin 0.950 versus 0.963 for pattern recognition (difference = -0.013; 95% CI = -0.034 to 0.005). Using the cut-off 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 (28, 29). 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 UK the Royal College of Obstetricians and Gynaecologists (RCOG) recommends the use of the RMI (30, 31), but in (9) we compared the performance of the two IOTA logistic regression models with the RMI on a test set of 236 patients and demonstrated that the RMI performed significantly worse: AUC 0.936 and 0.916 for the logistic regression models versus 0.870 for RMI ($p = 0.0038$) (9). The strategy followed by the American College (ACOG) is the use of guidelines based on demographic and ultrasound variables (CA 125 level: > 35 U/ml for postmenopausal patients and > 200 U/ml for premenopausal patients, evidence of ascites on ultrasound or CT, a nodular or fixed pelvic mass, evidence of abdominal or distant metastases on CT, 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 centre with a gynaecologic oncology department. (32).

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

Since the performance of these "gold RCOG and ACOG 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 can not compare the performance of the IOTA models with that of the RCOG or ACOG guidelines because we have not prospectively collected information on physical examination or CT 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.

The robustness of the IOTA models may be explained by them having been developed on a large dataset where data had been collected from many centres and using a standardized ultrasound protocol and standardized terms and definitions. On the other hand, the models were validated in the three centres that had provided most of the data to the IOTA phase I 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 cut-off 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, i.e., 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.

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. Even though 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, i.e., they need to be tested in completely new centers that did not take part in the IOTA phase I 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.

ACKNOWLEDGEMENTS

This research was supported by the research council of the Katholieke Universiteit Leuven, Belgium (GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering OPTEC); the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011); the EU: BIOPATTERN (FP6-2002-IST 508803); ETUMOUR (FP6-2002-LIFESCIHEALTH 503094); Healthagents (IST–2004–27214); the Swedish Medical Research Council (grants nos. K2001-72X-11605-06A, K2002-72X-11605-07B, K2004-73X-11605-09A and K2006-73X-11605-11-3); funds administered by Malmö University Hospital; Allmänna Sjukhusets i Malmö Stiftelse för bekämpande av cancer (the Malmö General Hospital Foundation for fighting against cancer); ALF-medel and Landstingsfinansierad regional forskning (two Swedish governmental grants). The funding agencies had no role in the design of the study, the collection of the data, the analysis and interpretation of the data, the writing of the manuscript or the decision to submit the manuscript for publication.

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: 11-16

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: 338–347

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: 192-197

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. Gyn Oncol 2001; 80: 162-167

8. Van Holsbeke C, Van Calster B, Valentin L, Testa AC, Ferrazzi E, Dimou I, Lu C, Moerman P, Van Huffel S, Vergote I, Timmerman D. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the international ovarian tumor analysis 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: 500-505

11. Van Calster B, Timmerman D, Lu C, Suykens J, Valentin L, Van Holsbeke C, Amant F, Vergote I, Van Huffel S. Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. UOG 2007; 29: 496-504

12. Van Calster B, Timmerman D, Nabney I, Valentin L, Testa AC5, Van Holsbeke C, Vergote I, Van Huffel S. Using Bayesian Neural Networks with ARD input selection to detect malignant adnexal masses prior to surgery 2007; Neural computing and applications. In press.

13. Use of morphology to characterize and manage common adnexal masses. Valentin L, Best Pract Res Clin Obs Gyn 2004; 18: 71-89

14. Lu C., Suykens J.A.K., 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, (Ichimura T. and Yoshida K. , ed.), vol. 7 of International Series on Advanced Intelligence, Advanced Knowledge International (Magill, Australia), 2004, 343-382

15. Tipping ME. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 2001; 1:211-244

16. Pelckmans K, Goethals I, De Brabanter J, Suykens JAK, De Moor B (2005). Componentwise least squares support vector machines. In L Wang (ed), Support vector machines: Theory and applications. Berlin: Springer, pp 77-98

17. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J (2002) Least squares support vector machines. Singapore: World Scientific

18. Pochet N, Suykens J. Support vector machines versus logistic regression: improving prospective performance in clinical decision-making. Ultrasound Obstet Gynecol 2006; 7: 607-608

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

20. Nabney IT (2002) NETLAB. Algorithms for pattern recognition. Springer.

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: 190-195

23. Pepe M, Longton G, Anderson G, Schummer M. Selecting differentially expressed genes from micorarray experiments. Biometrics 2003; 59: 133-142

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. Statistics in Medicine 1998; 17: 2635-2650

28. Valentin L. Use of morphology to characterize and manage common adnexal masses. Best Pract Res Clin Obstet Gynaecol. 2004; 18: 71-89

29. Valentin L. Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. Ultrasound Obstet Gynecol 1999; 14: 338-347

30. 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: 922-929

31. RCOG guideline no. 34, October 2003

32. 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

Table 1: Description of all 11 models developed on the IOTA phase 1 dataset.

	Type of model	Cut-off	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 50mm), (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, (6) acoustic shadows
Bay MLP 11-2a (12)	Artificial neural network	0.15	(1) age, (2) hormonal therapy, (3) ascites , (4) max diameter of the lesion, (5) irregular internal cyst wall, (6) color score, (7) blood flow within papillary projection, (8) number of papillary projection, (9) max diameter of solid component, (10) multilocular-solid tumor, (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) max diameter of the lesion, (6) irregular internal cyst wall, (7) blood flow within papillary projection, (8) acoustic shadowing, (9) max diameter of solid component, (10) unilocular tumor, (11) solid tumor
Bay Perc 11 (12)	Artificial neural network	0.15	(1) age, (2) personal history of ovarian cancer, (3) ascites, (4) max diameter of the lesion, (5) irregular internal cyst wall, (6) color score, (7) blood flow within papillary projection, (8) acoustic shadowing, (9) max diameter of solid component, (10) unilocular tumor, (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)	Relevance vector machine	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)	Relevance 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 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)	Relevance 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 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

LR: logistic regression model; LS-SVM: least squares support vector machine; lin: linear; rbf: radial basis function; add rbf: additive radial basis function; RVM: relevance vector machine: Bay: Bayesian; MLP: multi-layer perceptron; Perc: perceptron

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

	All (n = 507)		Leuven (n = 287)		Rome (n = 96)		Malmö (n = 124)	
Histology	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
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
Prim 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*

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		Theta
Variable	**N**	**Median**	**N**	**Median**	**(95% CI)**
Age, yrs	202	40	85	58	0.77 (0.70-0.82)
Max Diam Lesion, mm	202	52	85	89	0.70 (0.63-0.76)
Nr of 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)
CA125, U/mL	109	17	81	296	0.88 (0.82-0.92)
Rome new (n = 96)	**Benign**		**Malignant**		**Theta**
Variable	**N**	**Median**	**N**	**Median**	**(95% CI)**
Age, yrs	64	45	32	52	0.64 (0.51-0.74)
Max Diam Lesion, mm	64	78	32	101	0.64 (0.51-0.74)
Nr of 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)
CA125, U/mL	20	38.5	22	220	0.75 (0.57-0.86)
Malmö new (n = 124)	**Benign**		**Malignant**		**Theta**
Variable	**N**	**Median**	**N**	**Median**	**(95% CI)**
Age, yrs	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)
Nr of 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)
CA125, U/mL	88	18	25	251	0.86 (0.75-0.92)

Max Diam Lesion, largest diameter of the lesion; Max Diam Solid, largest diameter of the largest solid component of the tumor; diam, diameter; CI, Confidence Interval

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		Diff in %
Variable	N	% yes	N	% yes	(95% CI)
Pers Hist Ov Ca	202	2 (1.0)	85	4 (4.7)	-3.7 (-10.5, 0.1)
Hormonal therapy	202	38 (18.8)	85	17 (20.0)	-1.2 (-12.0, 8.1)
Pelvic pain	202	71 (35.1)	85	20 (23.5)	11.6 (-0.2, 21.9)
Ascites	202	3 (1.5)	85	47 (55.3)	-53.8 (-64.0, -42.9)
Pap blood flow	37	7 (18.9)	49	43 (87.8)	-68.9 (-80.3, -49.4)
Internal wall, irregular	202	93 (46.0)	85	62 (72.9)	-26.9 (-37.6, -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, -20.9)
Entirely solid tumor	202	30 (14.9)	85	29 (34.1)	19.2 (-30.7, -8.5)
Rome new (n = 96)	**Benign**		**Malignant**		**Diff in %**
Variable	**N**	**% yes**	**N**	**% yes**	**(95% CI)**
Pers Hist Ov Ca	64	0 (0.0)	32	0 (0.0)	0.0 (-10.7, 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, 12.7)
Ascites	64	1 (1.6)	32	5 (15.6)	-14.0 (-30.2, -3.0)
Pap blood flow	14	4 (28.6)	9	7 (77.8)	-49.2 (-72.4, -7.5)
Internal wall, irregular	64	26 (40.6)	32	25 (78.1)	-37.5 (-53.1, -16.7)
Acoustic shadows	64	9 (14.1)	32	1 (3.1)	10.9 (-3.3, 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, -1.2)
Entirely solid tumor	64	7 (10.9)	32	16 (50.0)	-39.1 (-56.3, -19.9)
Malmö new (n = 124)	**Benign**		**Malignant**		**Diff in %**
Variable	**N**	**% yes**	**N**	**% yes**	**(95% CI)**
Pers Hist Ov Ca	98	1 (1.0)	26	0 (0.0)	1.0 (-11.9, 5.6)
Hormonal therapy	98	30 (30.6)	26	3 (11.5)	19.1 (-0.2, 31.4)
Pelvic pain	98	8 (8.2)	26	1 (3.8)	4.4 (-11.2, 12.1)
Ascites	98	0 (0.0)	26	2 (7.7)	-7.7 (-24.1, -1.0)
Pap blood flow	19	8 (42.1)	11	11 (100)	-57.9 (-76.9, -24.2)
Internal wall, irregular	98	29 (29.6)	26	21 (80.8)	-51.2 (-64.6, -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, -0.9)
Entirely solid tumor	98	4 (4.1)	26	10 (38.5)	-34.4 (-53.6, -17.3)

Pers Hist Ov Ca, personal history of ovarian cancer; 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 I dataset when tested on the test set of the International Tumor analysis (IOTA) phase I dataset and when tested prospectively on the current study population. Because of the ranking analysis, the models are ranked from high to low AUC for all three centers combined.

	AUC performance			Cut-off performance				Ranking analysis		
	AUC IOTA 1 test	AUC ALL	AUC	Sens	Spec	LR+	LR-	Mean rank	Pm(1)	Pm(3)
Model	set (95% CI) n = 312	(95% CI) n = 507	Leave-Rome-Malmö n = 287-96-124							
LS-SVM lin (11)	.946 (.911–.973)	.950 (.931–.967)	.952–.889–.986	91.6	83.8	5.65	0.10	3.7	.148	.485
LR 1 (9)	.942 (.905–.972)	.950 (.929–.967)	.956-.903–.992	95.1	74.2	3.68	0.07	4.4	.145	.472
LR 2 (9)	.922 (.882–.954)	.950 (.928–.967)	.957–.884–.983	95.1	73.9	3.64	0.07	4.8	.209	.423
LS-SVM rbf (11)	.945 (.909–.971)	.949 (.930–.966)	.952–.891–.981	91.6	83.2	5.47	0.10	4.6	.048	.339
Bay MLP 11-2a (12)	.942 (.900–.974)	.948 (.928–.964)	.949–.912–.985	93.7	76.9	4.06	0.08	5.9	.150	.303
LS-SVM add rbf (11)	.943 (.905–.970)	.948 (.929–.963)	.954–.885–.980	93.7	81.9	5.17	0.08	6.3	.082	.204
RVM lin (11)	.949 (.913–.975)	.948 (.926–.965)	.949–.892–.984	91.6	82.1	5.13	0.10	6.6	.024	.137
Bay MLP 11-2b (12)	.933 (.892–.964)	.947 (.926–.966)	.958–.887–.975	95.1	78.8	4.50	0.06	6.4	.118	.293
RVM add rbf (11)	.946 (.910–.973)	.947 (.927–.964)	.954–.878–.980	93.7	81.3	5.02	0.08	7.0	.026	.128
Bay Perc 11 (12)	.947 (.908–.974)	.946 (.923–.963)	.953–.905–.982	95.8	78.3	4.41	0.05	7.7	.045	.186
RVM rbf (11)	.946 (.910–.972)	.945 (.923–.964)	.947–.892–.981	92.3	80.2	4.67	0.10	8.6	.001	.022
Pattern recognition	.963			90.2	92.9	12.7	0.11			

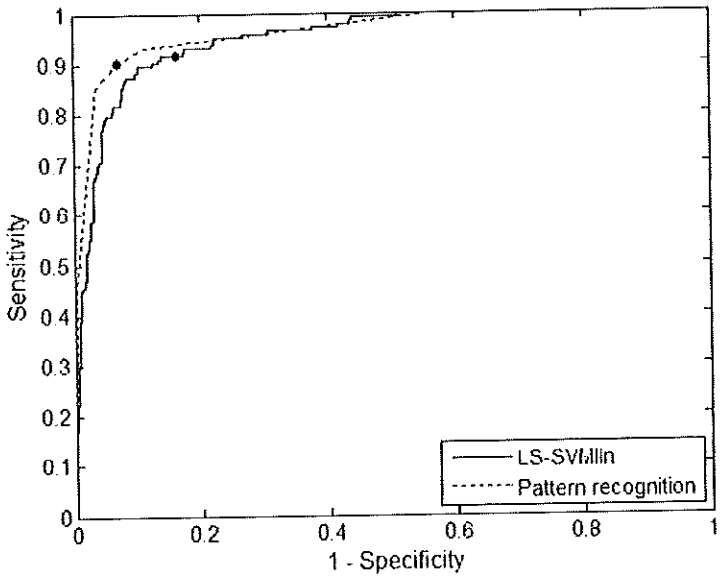
AUC, area under the receiver operating characteristic curve; CI, confidence interval; Sens, sensitivity; Spec, specificity; LR+,positive likelihood ratio; LR-, negative likelihood ratio; LR, logistic regression model; LS-SVM, least squares-support vector machine; lin, linear; rbf, radial basis function; add rbf, additive radial basis function; MLP, multi-layer perceptron; Bay, Bayesian; Perc, perceptron; RVM, relevance vector machine.
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.

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

n = 507	AUC	Cut-off	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; 0.005		difference = 1.4%, 95% CI = -4.2% to 7.1%	difference = -9.1%, 95% CI = -13.1% to -5.2%		

LR+: positive likelihood ratio; LR -: negative likelihood ratio; LS-SVM lin: least squares support vector machine with linear kernel; *: 95% CI was obtained using the bias-corrected bootstrap method (24).

FIGURE 1. Receiver operating characteristic curve for the best IOTA model LS-SVM lin (11) and pattern recognition (13) based on the level of diagnostic confidence. The performance when using the cut-offs (0.15 for LS-SVM lin, prediction of malignancy for pattern recognition) is indicated with dots.


Sensitivity
1 - Specificity
LS-SVMlin
Pattern recognition
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8