OC150
Investigation of the performance of mathematical models on small ovarian masses in the IOTA phase 1 and 2 study

O. Gevaert[1], A. C. Testa[2], A. Daemen[1], C. Van Holsbeke[3], R. Fruscio[4], E. Epstein[5], F. P. G. Leone[6], A. Czekierdowski[7], L. Valentin[8], L. Savelli[9], T. Bourne[10], F. Amant[11], B. De Moor[1], D. Timmerman[11]

[1]Dept Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Leuven, Belgium, [2]Università Cattolica del Sacro Cuore, Rome, Italy, [3]University Hospitals, London, Belgium, [4]San Gerardo Hospital, Monza, Italy, [5]Lund University Hospital, Lund, Sweden, [6]DSC L. Sacco Università di Milano, Milano, Italy, [7]Medical University, Lublin, Poland, [8]University Hospital, Malmo, Sweden, [9]Reproductive Medicine Unit, Bologna, Italy, [10]St Georges Hospital Medical School, London, United Kingdom, [11]University Hospitals, Leuven, Belgium

Objectives: The first phase of IOTA resulted in a data set of 1066 patients from 9 centers in 5 countries. Previously, this data set was randomly stratified in 70% of patient data to construct a logistic regression model (referred to as M1) and 30% of the patient data as a test set to estimate the predictive performance. IOTA phase 2 resulted in a data set of 1940 patients from 19 centers in 8 countries. We investigate whether the performance of M1 depends on the size of ovarian masses when used prospectively on the IOTA phase 2 data set.
Methods: The performance of M1 was estimated on the IOTA phase 2 data set by calculating the Area Under the ROC curve (AUC) on patients with a maximum lesion diameter smaller than a predefined threshold. This threshold was then iteratively increased to investigate the evolution of the AUC as a function of the size of the ovarian mass.
Results: We observed a significant decrease of the AUC on the IOTA phase 2 data set when the maximum diameter of the lesion is increased from 28 mm to 32 mm. The AUC for all masses with a maximum lesion diameter smaller than 29 mm is 0.947 (SE 0.023) while the AUC for all masses with a maximum lesion diameter smaller than 33 mm is 0.889 (SE 0.047). When focusing on this subgroup of patients with ovarian masses with a maximum lesion diameter from 29 to 32 mm, we found 61 patients which were significantly younger (P-value = 0.057) and had a different color score distribution (P-value 0.0066) compared to the remaining patients from the IOTA 2 data set. There were 6 malignant masses (10%) in this set of patients, while M1 predicted 14 masses to be malignant (4 correct) when using the previously determined threshold of 0.1 for classifying ovarian masses. Similar results were observed on the smaller IOTA phase 1 test set.
Conclusions: These results indicate that masses with a maximum lesion diameter from 29 till 32 mm are hard to classify for mathematical models.

OC151
Pattern recognition by less experienced examiners and use of mathematical models to discriminate between static ultrasound images of benign and malignant adnexal masses

C. Van Holsbeke[1], L. Lannoo[2], T. Mesens[1], E. de Jonge[1], L. Valentin[3], D. Jurkovic[4], J. Yazbek[4], T. Holland[4], D. Timmerman[2], A. Daemen[5]

[1]Ziekenhuis Oost-Limburg, Genk, Belgium, [2]University Hospitals Leuven, Leuven, Belgium, [3]Department of Obstetrics and Gynecology, Malmö University Hospital, Lund University, Malmö, Sweden, [4]Department of Obstetrics and Gynaecology, King's College Hospital, London, United Kingdom, [5]Department of Electrical Engineering, ESAT-SCD, Catholic University, Leuven, Belgium

Aim: To evaluate how accurate less experienced sonologists can classify adnexal masses when using pattern recognition or mathematical models.
Methods: Static images from an artificial collection of adnexal masses were evaluated by two senior registrars before and after an extra training in gynecological ultrasound. They had to classify the masses as benign or malignant using pattern recognition, the main IOTA logistic regression model and the IOTA scoring system.
Results: 165 masses were examined of which 58% were benign and 42% malignant on histology; 49% of the malignant masses were borderline tumors. After training, pattern recognition by the two examiners reached a sensitivity of 70% and 61% and a specificity of 92% and 95%.
Training decrease sensitivity ($P = 0.0039$ and 0.0578) and increased specificity ($P = 0.001$ and 0.0578).
When the scoring system was assessed, the sensitivity was 59% and 54% and the specificity 90% and 93%.
For the main logistic regression model sensitivity was 70% and 56% and specificity 84% and 94%.
The main reasons for the misclassification of malignant adnexal masses were: failure to recognize solid components or papillary projections, failure to appreciate irregularity of the cyst wall, incorrect interpretation of acoustic shadowing, and omit to include the color score or personal history of ovarian cancer.
Conclusions: Whatever strategy was used by the less experienced sonologist, specificity was very high but sensitivity was disappointing. The main aim of developing mathematical models to discriminate between benign and malignant adnexal masses is to help less experienced examiners. Despite the fact that this study used an artificial collection of difficult masses and that the examiners could only evaluate static images, it shows that before using any kind of model, one should be able to assess an adnexal mass and to recognize the most important features. Training should focus on recognizing features that are typical for malignant tumors.

OC152
Prevalence of cancer and optimal cut-off levels for mathematical models to distinguish between benign and malignant adnexal masses

A. Daemen[1], C. Van Holsbeke[2], R. Fruscio[3], S. Guerriero[4], A. Czekierdowski[5], L. Valentin[6], L. Savelli[7], A. C. Testa[8], N. Colombo[9], T. Bourne[10], I. Vergote[11], B. De Moor[1], D. Timmerman[11]

[1]Department Elektrotechniek - ESAT/SISTA, Katholieke Universiteit Leuven, Leuven, Belgium, [2]Ziekenhuis Oost-Limburg, Genk, Belgium, [3]San Gerardo Hospital, Monza, Italy, [4]Ospedale San Giovanni di Dio, Cagliari, Italy, [5]Medical University, Lublin, Poland, [6]University Hospital, Malmo, Sweden, [7]Reproductive Medicine Unit, Bologna, Italy, [8]Universita Cattolica del Sacro Cuore, Rome, Italy, [9]Prof. ssa Ginecologic Oncology Unit, IEO, Milano, Italy, [10]St Georges Hospital Medical School, London, United Kingdom, [11]Department of Obstetrics and Gynecology, Katholieke Universiteit Leuven, Leuven, Belgium

Objectives: Two logistic regression models LR1 and LR2 to distinguish between benign and malignant adnexal masses were developed in phase 1 of a multicenter study by the International Ovarian Tumor Analysis (IOTA) group. The goal of this retrospective analysis is to verify if the models perform differently between types of center and if the cut-off levels of the models require alteration per center or type of center.
Methods: 19 centers participated in this study and contributed 1940 new cases. Concerning the types, a distinction is made according to the prevalence of malignant cases into centers with

a prevalence of less than 15%, between 15 and 30% and above 30%. To ascertain statistically significant differences in performance between the types of centers, the AUCs were compared using bootstrapping. The optimal cut-off level per center and type was chosen corresponding to a sensitivity level as high as possible (preferable above 90%) while still keeping a good specificity (80%) as this was considered to be very important in correctly identifying malignant cases.
Results: Both LR1 and LR2 performed better, although not statistically significant, in centers with a lower prevalence of malignant cases. The AUC of centers with less than 15% of malignancy was 0.956 and 0.941, for LR1 and LR2 respectively; centers with prevalence between 15 and 30% had an AUC of 0.948 and 0.925, respectively and centers with more than 30% malignancies had an AUC of 0.933 and 0.914, respectively. The optimal cut-off per center varied between 0.05 and 0.20, but the performance in the centers with a higher percentage of malignant cases did not improve by choosing a different cut-off level.
Conclusions: The performance of the logistic regression models increases with decreasing prevalence of malignancy. Because new cut-off levels per center would be based on 8 to 253 patients and the cut-off of 0.10 is optimal for all three types of center, it seems reasonable to use this cut-off in all centres.

OC153
Prospective evaluation of a model to diagnose adnexal masses as benign, primary invasive, borderline, or metastatic

B. Van Calster[1], C. Van Holsbeke[2], R. Fruscio[3], S. Guerriero[4], A. Czekierdowski[5], L. Valentin[6], L. Savelli[7], A. C. Testa[8], D. Paladini[9], F. P. G. Leone[10], E. Epstein[11], T. Bourne[12,13], S. Van Huffel[1], D. Timmerman[13]

[1]Katholieke Universiteit Leuven, Leuven, Belgium, [2]UZ Leuven and Ziekenhuis Oost-Limburg, Leuven and Genk, Belgium, [3]San Gerardo Hospital, Monza, Italy, [4]Ospedale San Giovanni di Dio, Cagliari, Italy, [5]Medical University, Lublin, Poland, [6]University Hospital, Malmö, Sweden, [7]University of Bologna, Bologna, Italy, [8]Università Cattolica del Sacro Cuore, Rome, Italy, [9]Università degli Studi di Napoli "Federico II", Naples, Italy, [10]Università di Milano, Milan, Italy, [11]University Hospital, Lund, Sweden, [12]St George's Hospital, London, United Kingdom, [13]UZ Leuven, Leuven, Belgium

Objectives: To develop and prospectively evaluate a diagnostic model that differentiates between benign, primary invasive, borderline, and metastatic adnexal masses.
Methods: Using various mathematical techniques, diagnostic models for the differentiation between four types of adnexal masses (benign, primary invasive, borderline and metastatic malignancies) were developed. Model development was carried out using the IOTA (International Ovarian Tumor Analysis group) phase 1 multi-center data set ($n = 1066$). The best model as assessed on the test part of this data was prospectively evaluated on the IOTA phase 2 data ($n = 1940$). Model performance was evaluated using receiver operating characteristic curves and the area under them (AUC).
Results: The best model was a combination of six binary logistic regression models, with the binary models contrasting pairs of tumor types (e.g. benign vs borderline or metastatic vs borderline). The IOTA phase 2 data set was collected at 7 centers from phase 1 and 12 new centers. There were 1395 benign, 375 primary invasive (including rare malignancies), 112 borderline, and 58 metastatic masses. The AUCs for these tumor types were 0.944, 0.925, 0.877, and 0.833, respectively. Using only the patients from new centers, the AUCs were 0.950, 0.935, 0.909, and 0.843, respectively. Based on the training set results, we classified a tumor as benign if the probability of a benign tumor was at least 0.84. The other tumors were classified as borderline if the probability of a borderline tumor was at least 0.12. All remaining tumors were classified as invasive (primary or metastatic invasive). Using this system, 84% of benign, 58% of borderline, and 74% of invasive masses were correctly classified. In general, 91% of malignant masses were given a malignant prediction.
Conclusions: The model to distinguish between four types of adnexal masses had good performance on prospectively collected data. Even for borderline and metastatic tumors, AUCs were high.

OC154
Prospective external validation of an ultrasound scoring system to differentiate benign from malignant masses in specific subgroups of adnexal tumors

L. Valentin[1], L. Ameye[2], R. Fruscio[3], C. Van Holsbeke[4], A. Czekierdowski[5], S. Guerriero[6], N. Colombo[7], D. Fischerova[8], B.C.H. Jingzhang[9], S. Van Huffel[2], A. C. Testa[10], D. Timmerman[11]

[1]Malmö University Hospital, Lund University, Malmö, Sweden, [2]Department of Electrical Engineering, ESAT-SISTA, Katholieke Universiteit Leuven, Leuven, Belgium, [3]San Gerardo Hospital, Monza, Italy, [4]Ziekenhuis Oost-Limburg, Genk, Belgium, [5]Medical University in Lublin, Lublin, Poland, [6]University of Cagliari, Cagliari, Italy, [7]European Institute of Oncology, Milano, Italy, [8]First Medical Faculty of Charles University in Prague, Prague, Czech Republic, [9]Chinese PLA General Hospital, Beijing, China, [10]Università Cattolica del Sacro Cuore, Rome, Italy, [11]University Hospitals KU Leuven, Leuven, Belgium

Objectives: To determine the diagnostic performance of an ultrasound scoring system to differentiate benign from malignant adnexal tumors and to compare its performance to that of subjective evaluation of ultrasound findings, i.e., pattern recognition.
Methods: In the International Ovarian Tumor Analysis (IOTA) study phase 1 ($n = 1066$) an ultrasound scoring system was developed for discrimination between benign and malignant tumors in four subgroups of tumor (unilocular cyst, multilocular cyst, presence of a solid component but no papillary projections, and presence of papillary projections). This scoring system was tested prospectively in the IOTA study phase 2. This study includes 1940 patients with a persistent adnexal mass, who underwent transvaginal gray scale and color Doppler ultrasound examination by experienced examiners using good ultrasound equipment and a standardized examination technique, terms and definitions. The sensitivity and specificity with regard to malignancy of the scoring system was compared with those of pattern recognition.
Results: There were 545 (28%) malignancies. The sensitivity of the scoring system was 73% (396/545) and that of pattern recognition was 90% (492/545) ($P < 0.001$). The specificity was 93% (1298/1395) for the scoring system and 93% (1292/1395) for pattern recognition ($P = 0.55$).
Conclusions: The diagnostic performance of the subgroup scoring system was poorer than that of pattern recognition. This is explained by its much lower sensitivity. We plan to fine tune the scoring system in a larger study population.

OC155
Normative data of the transverse diameter of the developing fetal thymus

F. Gamez[1], J. Santolaya-Forgas[2], J. A. De Leon[1], P. Pintado[1], R. Perez-Fernandez[1], L. Ortiz-Quintana[1]

[1]Hospital General Gregorio Maranon, Madrid, Spain, [2]Brigham and Women's Hospital, Boston, United States