This is an Accepted Manuscript of an article published by Elsevier in Gynecologic Oncology on 2016-06-30, available online: https://dx.doi.org/10.1016/j.ygyno.2016.06.020

External validation of the IOTA ADNEX model performed by two independent gynecologic centers.

Sebastian Szubert[1], Andrzej Wojtowicz[2], Rafal Moszynski[1], Patryk Zywica[2], Krzysztof Dyczkowski[2], Anna Stachowiak[2], Stefan Sajdak[1], Dariusz Szpurek[1], Juan Luis Alcazar[3]

[1] Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland

[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznan, Poland

[3] Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain

Key words: adnexal mass; ovarian tumor; ovarian cancer; ultrasound; ADNEX model

Corresponding author:

Sebastian Szubert, M.D., Ph.D.

Division of Gynecological Surgery, 33. Polna St.; 60-535 Poznan, Poland,

phone +48618419490 fax +48 61 8419418

E-mail: szuberts@o2.pl

Abstract

Objectives: The external, two-center validation of the IOTA ADNEX model for differential diagnosis of adnexal tumors. **Methods:** A total of 204 patients with adnexal masses (134 benign and 70 malignant) treated at the Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland (Center I), and 123 patients (89 benign and 34 malignant) from the Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain (Center II), were enrolled into the study. **Results:** ADNEX achieved high accuracy in discriminating between malignant and benign ovarian tumors in both centers (79.9% and 81.3% in Centers I and II, respectively). Multiclass accuracy was substantially lower than in binary classification (malignant vs. benign): 64.2% and 74.0% in Centers I and II, respectively. Sensitivity and specificity for the diagnosis of specific tumor types in Center I were as follows: benign tumors– 72.4% and 94.3%; borderline tumors – 33.3% and 87.0%, stage I ovarian cancers – 00.0% and 91.8%; stage II – IV ovarian cancers – 68.2% and 83.1%; and metastatic tumors – 00.0% and 99.5%. Sensitivity and specificity in Center II were as follows: benign tumors – 75.3% and 97.1%; borderline tumors – 50.0% and 88.2%, stage I ovarian cancers – 40.0% and 97.5%; stage II – IV ovarian cancers – 95.0% and 88.3%; and metastatic tumors – 20.0% and 98.3%. **Conclusions:** ADNEX is characterized by very high accuracy in differentiating between malignant and benign adnexal tumors. However, prediction of ovarian tumor types could be more accurate.

Introduction

Differential diagnosis of ovarian tumors remains one of the most challenging problems in modern gynecology. Adequate discrimination between malignant and benign ovarian tumors determines the need for surgery, venue, surgical skills (a gynecologic oncologist or a general gynecologist), and operative time[1]. The International Ovarian Tumor Analysis (IOTA) group, after the publication of terms and definitions used in the description of sonographic images, started a new era in ultrasonographic diagnosis of adnexal masses[2]. The great effort of the IOTA group resulted in the development of new mathematical models characterized by high diagnostic performance [3]. In a meta-analysis by Kaijser et al., logistic regression model (LR2) and simple rules (SR) achieved higher diagnostic accuracy than other studied models, including risk of malignancy indices (RMI) [4].

The ADNEX model is one of the recent developments of the IOTA group. Apart from high accuracy in differentiating between malignant and benign adnexal masses, ADNEX allows for a more specific diagnosis of the tumor. Additionally, the model classifies malignant adnexal masses into borderline, stage I invasive, stage II-IV invasive, and secondary metastatic tumors[5]. Preoperative information about the subtype of the malignant tumor is essential for surgery planning. Early-stage ovarian malignancies require less invasive surgery as compared to advanced-stage disease. Thus, operative time may be shortened. Borderline tumors, and even some early-stage ovarian cancers, may be treated with fertility-sparing surgery[6,7]. Furthermore, suspicion of metastatic ovarian tumor requires careful preoperative search for the origin of cancer type. Taking all these factors into account, we believe that an accurate

diagnostic tool, which allows for a subdivision of malignant ovarian tumors, will be useful in clinical practice.

The IOTA group has already demonstrated high accuracy of ADNEX for both, development and validation data[5]. However, new studies are needed for adequate characterization of ADNEX applicability. Thus, the main aim of the presented study was external validation of the ADNEX model, performed by gynecologists from two independent European centers.

Material and methods

Data collected in ultrasonographic database of ovarian tumors in patients referred to our clinics for the application of the ADNEX model were retrospectively evaluated. In case of the Division of Gynecologic Surgery, Poznan University of Medical Sciences (Center I), data used in the study were obtained from patients treated due to ovarian tumors between December 2012 and April 2015. The study included only patients evaluated by R.M. and S.Sz. on days 1-3 before surgery. Only patients with complete data required for the ADNEX calculation were enrolled. There were no specific exclusion criteria and the only inclusion criterion was the need for surgery due to ovarian tumor. The tumors were evaluated using Aloka Alpha 10 with 3.75 – 7.5 MHz endovaginal probe and Aloka 3500 with a 7.5 MHz endovaginal probe (Hitach Aloka, Tokyo, Japan). A transabdominal probe was used in case of large tumors. In case of bilateral ovarian tumors, data on tumors with more complex morphology were collected. The analyzed tumors were ultrasonographically assessed according to the 2000 IOTA criteria[2]. CA125 serum levels were assessed on days 1-5 before surgery using an immunoenzymatic test (ST AIA-PACK OVCATosoH Bioscience, Tokyo, Japan).

R.M. has over 12years of experience in gynecological ultrasonography (approximately 800 examinations each year). S.Sz. has 8 years of experience in gynecological ultrasonography (approximately 1000 examinations each year). Both, R.M. and S.Sz. conduct clinical studies concerning gynecological ultrasonography and participate in numerous courses and lectures about ultrasound examination. Both, R.M and S.Sz. can be classified as level 2 examiners, according to the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) criteria.

Data collected from January 2011 to October 2012 at the Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain, (Center 2) were also included into the study. The study included patients treated due to ovarian tumor, with complete clinical and ultrasonographic characteristic required for the ADNEX calculation. There were no specific exclusion criteria, and the only inclusion criterion was the need for surgery due to ovarian tumor. All women underwent transvaginal or transrectal ultrasound using a Voluson E8 equipped with an RIC5-9MHz endovaginal probe (GE Healthcare, Milwaukee, USA). In case of a large tumor transabdominal ultrasound was also performed. All examinations were performed by one physician (J.L.A) using the standards and terminology proposed by the IOTA group [2]. CA125 serum levels were evaluated on days 1-5 before ultrasonographic evaluation using Cobas-Core CA-125-II, Laboratories Roche, Basel, Switzerland.

J.L.A has over 20 years of experience in gynecological ultrasonography. He is the author of numerous studies concerning differential diagnosis of ovarian tumors. J.L.A is a teacher of gynecological ultrasonography and he is regarded as an expert in this field. J.L.A. fulfils all criteria of being level 3 practitioner according to the EFSUMB criteria.

Following the ultrasound examination, the observers from both Centers made their suppositions about tumor character, classifying the mass either as malignant or benign. The obtained results were used to evaluate the accuracy of the subjective ultrasonographic assessment. The observers were blinded to the results of the ADNEX calculation.

All tumors diagnosed and treated in both gynecologic centers were surgically removed. Final histopathological diagnosis was obtained and the tumors were classified according to the WHO classification[8].

Statistical analysis was performed using R version 3.1.2 (2014-10-31) with libraries: 1) boot version 1.3.13 - to calculate bootstrap confidence intervals of performance measures, and 2) pROC version 1.8 - to calculate receiver operating characteristic (ROC), area under the ROC curve (AUC), and to compare AUCs. ADNEX performance was calculated using the formula obtained from the Appendix D of the manuscript by Calster et al.[5]. In case of 9 patients from Center 1, whose CA-125 levels were not available, we calculated the model using the formula provided on the ADNEX webpage (http://www.iotagroup.org/adnexmodel/). We used a cut-off of 10% to estimate the risk of malignancy. Both centers were treated as referral centers for gynecologic oncology patients. The performance measures were calculated for all patients and with reference to the menopausal status.

The comparison between AUC values from the two Centers was made using DeLong's test, while Fisher's Exact Test was used to compare subjective ultrasonographic assessment in the two Centers. The McNemar's test was performed to compare the ADNEX model performance with subjective assessment, considered as the referenced test for differential diagnosis of ovarian tumors. Statistical comparison of the

distributions of benign and malignant lesions between the two Centers was assessed using 2-sample test for equality of proportions with continuity correction. In the following study, in order to check the validity of the obtained results, we performed statistical tests and used resampling methods. We performed *t*-test power calculation to obtain the lower bound of sample size in testing differences between the expected AUC, as originally reported by ADNEX 0.943, and 0.5 (d = (0.943 − 0.5)/0.5), power equal to 0.99 and the significance level $\alpha = 0.05$. The estimated sample size was 22.

Results

Patient characteristics

Two hundred and four adnexal masses from ultrasonographic database of the Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland (Center 1), have data applicable for the ADNEX calculation. There were 134 (65.7%) benign and 70 (34.3%) malignant tumors. One hundred and thirty-eight (67.6%) and 66 (32.4%) tumors were diagnosed in pre- and postmenopausal women, respectively. Median patient age was 46 (range 15 – 84 years). Morphological classes of the investigated tumors were as follows: unilocular tumor – 47 (23%), unilocular solid tumor – 25 (12.3%), multilocular tumor 40 (19.6%), multilocular solid tumor– 68 (33.3%), solid tumor – 23 (11.3%), and 1tumor was regarded as not classifiable (0.5%). Twenty-two (10.8%) tumors had>10 locules, while 182 (89.2%) tumors had≤10 locules. The shadowing was present in 11 (5.4%) tumors, and free fluid in the pouch of Douglas was found in 42 (20.6%) patients. One hundred and sixty-six (81.4%) tumors had no papillary projections, while 4 (2.0%) had two, 7 (3.4%) had three, and 27 (13.2%)

tumors had more than three papillary projections. Median tumor maximal diameter was 80 mm (range 29 – 200 mm), while median solid part maximal diameter was 43mm (range 10 – 150 mm). Median CA125 level was 40 IU/ml (range 4 – 4909 IU/ml). CA125 was not available in 9 cases so ADNEX was calculated with the formula provided on the ADNEX webpage (http://www.iotagroup.org/adnexmodel/). The results of histopathological examinations are shown in Table 1.

Data obtained from 123 patients, diagnosed and treated due to adnexal masses at the Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain (Center II), between January 2011 and October 2012, were used to evaluate ADNEX applicability. The study group included 89 (72.4%) benign and 34 (27.6%) malignant adnexal masses. Seventy-two (58.5%) lesions were diagnosed in premenopausal, while 51 (41.5%) in postmenopausal women. Median patient age was 47 (range 12-81 years). There were 52 (42.3%) unilocular cysts, 4 (3.3%) unilocular solid tumors, 18 (14.6%) multilocular tumors, 13 (10.6%) multilocular solid tumors, and 36 (29.3%) purely solid tumors. Twenty tumors (16.3%) had >10 locules, and 103 (83.7%) tumors had ≤ 10 locules. The shadowing was observed in 18 (14.6%) tumors, and ascites was found in 22 (17.9%) patients. The occurrence of papillary projections was as follows: no papillary projections in 70 (56.9%), one papillary projection in 46 (37.4%), two papillary projections in 3 (2.4%), three papillary projections in 2 (1.6%), and more than three >3 in 2 (1.6%) tumors. Median tumor maximal diameter was 63 mm (range 16 – 262 mm) and median solid part maximal diameter was 39 mm (range 6– 132 mm). Median CA125 level was 40 IU/ml (range 1 – 3137 IU/ml). The results of the final postoperative histopathological examinations of the tumors included in the study are presented in Table 1.

Lesion distribution

The comparison of the proportion of benign and malignant lesions revealed no significant differences between Center I and II (P = 0.258). However, there were significant differences in tumor morphology between the two Centers (P < 0.001). The comparison of the distribution of ovarian tumor morphology between the two Centers revealed differences in the following subgroups: multilocular (P = 0.002), multilocular solid (P = 0.002), and solid tumors (P = 0.008), whereas the distribution of unilocular (P = 0.069), unilocular solid (P = 1.000), and not classifiable (P = 1.000) tumors was comparable.

ADNEX performance

ADNEX achieved high discrimination accuracy between malignant and benign ovarian tumors in both centers. The model had good diagnostic performance in the entirestudy population, as well as in premenopausal and postmenopausal subgroups. In the former, ADNEX resulted in very high AUC in Center II (0.955, 95% CI [0.893 – 0.996]), and high in Center I (0.907, 95% CI [0.858 – 0.948]). The differences between AUC were statistically insignificant (P = 0.171). In Center I, 66 (94.3%) malignant tumors were correctly classified, while 4 (5.7%) tumors were incorrectly classified as benign. On the other hand, 97 (72.3%) benign tumors were recognized properly, and there were 37 (27.6%) false positive results. In Center II, 33 (97.1%) of 34 tumors were classified correctly, and only 1 malignant tumor (2.9%) was misdiagnosed. There were 67 (75.3%) true positive and 22 (24.7%) false positive results in Center II.

The comparison of AUC between pre- and postmenopausal women indicated significantly higher AUC in the premenopausal group of patients in Center II (P = 0.046). However, there were no differences in AUC in Center I (0.974). Detailed

performance measurements for pre- and postmenopausal women from both centersare presented in Table 2. Figure 1 shows the ROC curves for the ADNEX model from both Centers.

Multiclass diagnostic accuracy for Center I (accuracy = 0.642, 95% CI [0.572 - 0.708]) and Center II (accuracy = 0.740 (95% CI [0.653 - 0.815]), was substantially lower than in binary classification (malignant vs. benign). The confusion matrix for multiclass classification for both centers is presented in Table 3. Table 4 includes diagnostic parameters for the analyzed sub-groups of adnexal masses. A comparison of the diagnostic performance of the ADNEX model for multi-classification revealed no statistically significant differences between Centers I and II. The Fisher's Exact Test with Bonferroni correction indicated the following p-values for the selected subclasses: benign tumors – 0.837, borderline tumors - 1.000, stage I ovarian malignancies - 0.070, stage II – IV malignancies – 0.100, and 1.000 for metastatic tumors.

Subjective ultrasonographic assessment performance

Sixty two (88.6%) out of 70 malignant adnexal masses, and 121 (90%) out of 134 benign tumors were correctly classified using subjective assessment in Center I. The overall diagnostic accuracy was 89.7%. The overall performance of subjective assessment was higher as compared to the ADNEX model ($P < 0.001$). Thirty three (97%) out of 34 malignant ovarian tumors were correctly classified as malignant tumors in Center II. Eighty four (94%) out of 89 were correctly diagnosed as benign tumors. The overall accuracy of subjective ultrasonographic assessment was 95.1% in Center II. Similarly, the diagnostic accuracy of subjective assessment was higher as compared to the ADNEX model ($P <0.001$). On the other hand, the difference in the performance of subjective ultrasonographic assessment between the two Centers was insignificant ($P =$

0.368). Detailed results of subjective ultrasonographic assessment for both Centers are summarized in Table 5.

Discussion

Our study has proven that sonographers from two independent European gynecologic centers could accurately discriminate malignant from non-malignant ovarian masses using the IOTAADNEX model. ADNEX is characterized by very high negative predictive value, thus it can be considered as an adequate tool for excluding malignancy. Our study provides high-level evidence for clinical applicability of the ADNEX model. In general, external validations of ultrasonographic models and indices present worse results than the original studies[9,10]. However, the results of ADNEX performance achieved in Center II were very similar to the original data provided by the IOTA group, and even better than those from validation data[5]. On the other hand, the results from Center I were substantially lower than those from Center II and the original paper by the IOTA group [5]. This may be explained by the differences in ultrasonographic experience between sonographers from Centers I (EFSUMB level 2) and II (EFSUMB level 3). Additionally, tumors evaluated in Center I presented with more complex morphology– 33.3% of multilocular-solid tumors and 13.2% of tumors had more than 3 papillary projections (compared to 10.6% and 1.6% in Center II, respectively).

Our results showed very high accuracy of the ADNEX model in distinguishing between malignant and non-malignant adnexal tumors, while the prediction of specific type of adnexal malignancy had lower performance. In our study, only 1 from 10 metastatic adnexal masses was correctly classified. On the other hand, the diagnosis of metastatic neoplasm was characterized by very high specificity. Our results are in somewhat

conflict with a study by Epstein et al.[11]. These authors showed high accuracy of the ADNEX model in diagnosing metastatic ovarian tumors – the model achieved 70% specificity at sensitivity near to 90%. This is of great importance because prediction of metastatic character of adnexal tumor is essential for surgery planning. Metastatic adnexal tumors often present specific ultrasonographic appearance, making it possible to predict them in subjective ultrasonographic assessment [12,13]. An experienced sonographer can even predict the site of origin of a metastatic adnexal tumor[12].However, the study by Epstein et al., compared subjective ultrasonographic assessment with the ADNEX model in prediction of a metastatic origin of adnexal tumor. Although not confirmed by statistical analysis, the ADNEX model seemed to be superior to subjective ultrasonographic assessment[11].

Our results indicate that the second weak point of the ADNEX model is its low sensitivity in diagnosing stage I ovarian malignancies. This is extremely important as its incidence is higher than metastatic lesions, and it may concern patients who select conservative treatment to preserve their fertility. However, similarly to diagnosing metastatic neoplasms, the detection of early-staged ovarian cancer was characterized by high specificity, because vast majority of them were correctly classify as malignant tumors. Early diagnosis of ovarian cancer has been a matter of concern for many years. Low sensitivity of ultrasonography-based techniques in the diagnosis of early-stage ovarian cancers is not surprising, because high-grade serous ovarian cancer – the most common type of ovarian malignancy – spread to advanced stage at a median diameter of about 3 cm [14]. The current strategy for early detection of ovarian cancer focuses on serial measures of CA125 levels [15] .

We found very similar results of ADNEX performance in both European gynecologic centers. The results were similar for the overall performance, as well as for the diagnosis of benign adnexal lesions, borderline tumors and metastatic adnexal neoplasms. Center II achieved better performance results for the diagnosis of early and advanced ovarian cancers. However, the differences were statistically insignificant, thus the discrimination between different types of tumors was consistent in both centers. One of the important differences between Centers I and II is that the latter found higher diagnostic accuracy in premenopausal patients as compared to their postmenopausal peers, contrary to the former. The original study by the IOTA group did not analyze ADNEX performance according to the menopausal status of the patients[5]. ADNEX doesnot incorporate the menopausal status of the patients –it focuses only on patient age. We had found previously that the menopausal status strongly influences the diagnostic applicability of predictive models used for differential diagnosis of ovarian tumours[16]. Thus, in our opinion, ADNEX performance with regard to patient menopausal status should be investigated.

In this study, we demonstrated the subjective ultrasonographic assessment to be superior to ovarian tumor evaluation using the ADNEX model. The subjective assessment was significantly better, irrespectively of examiner's experience, which is consistent with earlier reports about superiority of subjective ultrasonographic evaluation performed by an experienced ultrasonographer as compared to diagnostic models and indexes[17]. To the best of our knowledge, this has been the first comparison of the ADNEX performance with the subjective ultrasonographic assessment. Regardless, the results obtained using the ADNEX model were very satisfying, thus we speculate that, after further evaluation, this model could be used for

differential diagnosis of ovarian tumors when an experienced sonographer is not available.

Although it is challenging to compare diagnostic accuracy of two examiners on different dataset, the results of our study showed no significant differences in the ADNEX performance between the examiners with different skills (level 2 vs. level 3 according to EFSUMB). This was also reported by earlier studies, where IOTA models were shown to have very high diagnostic accuracy even in less experienced hands [18].

The main limitations of this study include its retrospective character and patient selection, which was non-consecutive. Both may be sources of bias. However, the ADNEX model requires specific data obtained after ultrasound examination performed by criteria strictly defined by the IOTA group [2]. All patients included in the study were evaluated according to these criteria, which were well-known to the examiners. Thus, we are of the opinion that in case of our Centers, a prospective study on consecutive patients will yield comparable results.

Conclusions

To the best of our knowledge, this has been the first external validation of the ADNEX model performed independently of the IOTA group. Our results indicate very high accuracy of ADNEX in terms of distinguishing between malignant and non-malignant ovarian masses. However, prediction of specific types of ovarian malignancy could be more accurate.

Conflict of interest statement

None of the authors have any conflicts of interest to disclose.

References:

[1] du Bois A, Rochon J, Pfisterer J, Hoskins WJ. Variations in institutional infrastructure, physician specialization and experience, and outcome in ovarian cancer: a systematic review. Gynecol Oncol 2009;112:422–36. doi:10.1016/j.ygyno.2008.09.036.

[2] Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol 2000;16:500–5. doi:10.1046/j.1469-0705.2000.00287.x.

[3] Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, et al. 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. doi:10.1200/JCO.2005.01.7632.

[4] Kaijser J, Sayasneh A, Van Hoorde K, Ghaem-Maghami S, Bourne T, Timmerman D, et al. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. Hum Reprod Update 20:449–62. doi:10.1093/humupd/dmt059.

[5] Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. BMJ 2014;349:g5920. doi:10.1136/bmj.g5920.

[6] Vasconcelos I, de Sousa Mendes M. Conservative surgery in ovarian borderline tumours: a meta-analysis with emphasis on recurrence risk. Eur J Cancer 2015;51:620–31. doi:10.1016/j.ejca.2015.01.004.

[7] Park J-Y, Kim D-Y, Suh D-S, Kim J-H, Kim Y-M, Kim Y-T, et al. Outcomes of fertility-sparing surgery for invasive epithelial ovarian cancer: oncologic safety and reproductive outcomes. Gynecol Oncol 2008;110:345–53. doi:10.1016/j.ygyno.2008.04.040.

[8] Tavassoli FA, Devilee P. Pathology and Genetics of Tumours of the Breast and Female Genital Organs. World Health Orgaanization Classification of Tumours. third. Lyon : International Agency for Research on Cancer, 2003; 2003.

[9] Van Holsbeke C, Van Calster B, Valentin L, Testa AC, Ferrazzi E, Dimou I, et al. External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the International Ovarian Tumor Analysis Group. Clin Cancer Res 2007;13:4440–7. doi:10.1158/1078-0432.CCR-06-2958.

[10] Mol BW, Boll D, De Kanter M, Heintz a P, Sijmons E a, Oei SG, et al. Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. Gynecol Oncol 2001;80:162–7. doi:10.1006/gyno.2000.6052.

[11] Epstein E, van Calster B, Timmerman D, Nikman S. Subjective ultrasound assessment, the ADNEX model and ultrasound guided tru-cut biopsy to differentiate disseminated primary ovarian cancer metastatic non-ovarian cancer. Ultrasound Obstet Gynecol 2015. doi:10.1002/uog.14892.

[12] Testa AC, Ferrandina G, Timmerman D, Savelli L, Ludovisi M, Van Holsbeke C, et al. Imaging in gynecological disease (1): ultrasound features of metastases in the ovaries differ depending on the origin of the primary tumor. Ultrasound Obstet Gynecol 2007;29:505–11. doi:10.1002/uog.4020.

[13] Testa AC, Mancari R, Di Legge A, Mascilini F, Salutari V, Scambia G, et al. The "lead vessel": a vascular ultrasound feature of metastasis in the ovaries. Ultrasound Obstet Gynecol 2008;31:218–21. doi:10.1002/uog.5251.

[14] Brown PO, Palmer C. The preclinical natural history of serous ovarian cancer: defining the target for early detection. PLoS Med 2009;6:e1000114. doi:10.1371/journal.pmed.1000114.

[15] Menon U, Ryan A, Kalsi J, Gentry-Maharaj A, Dawnay A, Habib M, et al. Risk Algorithm Using Serial Biomarker Measurements Doubles the Number of Screen-Detected Cancers Compared With a Single-Threshold Rule in the United Kingdom Collaborative Trial of Ovarian Cancer Screening. J Clin Oncol 2015;33:2062–71. doi:10.1200/JCO.2014.59.4945.

[16] Moszynski R, Zywica P, Wojtowicz A, Szubert S, Sajdak S, Stachowiak A, et al. Menopausal status strongly influences the utility of predictive models in differential diagnosis of ovarian tumors: an external validation of selected diagnostic tools. Ginekol Pol 2014;85:892–9.

[17] Meys EMJ, Kaijser J, Kruitwagen RFPM, Slangen BFM, Van Calster B, Aertgeerts B, et al. Subjective assessment versus ultrasound models to diagnose ovarian cancer: A systematic review and meta-analysis. Eur J Cancer 2016;58:17–29. doi:10.1016/j.ejca.2016.01.007.

[18] Sayasneh A, Wynants L, Preisler J, Kaijser J, Johnson S, Stalder C, et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. Br J Cancer 2013;108:2448–54. doi:10.1038/bjc.2013.224.

414 Table 1. Distribution of histopathological findings among tumors included in the study
415 depending on the Centre and patient menopausal status.

	premenopausal	postmenopausal	all
Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland (Centre I)			
Benign Brenner tumor	0	3	3
Cystadenofibroma	1	0	1
Adult teratoma	16	3	19
Corpus luteum cyst	2	0	2
Endometrioid cyst	43	2	45
Hemorrhagic cyst	4	0	4
Mucinous cystadenoma	9	6	15
Persistent ectopic pregnancy	1	0	1
Pedunculated leiomyoma	2	1	3
Serous cystadenoma	22	9	31
Theca cell tumor	1	3	4
Tubo-ovarian abscess	5	1	6
Serous borderline tumor	4	1	5
Mucinous borderline tumor	3	4	7
Clear cell adenocarcinoma	3	3	6
Endometrioid adenocarcinoma	4	1	5
Granulosa cell tumor	1	0	1
Metastatic ovarian tumor	3	2	5
Mucinous adenocarcinoma	1	1	2
Serous adenocarcinoma	7	18	25
Undifferentiated carcinoma	6	8	14
Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain (Centre II)			
Benign Brenner tumor	0	1	1
Cystoadenofibroma	1	2	3
Serous cystadenoma	4	6	10
Mucinous cystadenoma	5	10	15
Endometrioid cyst	37	3	40
Fibroma	2	4	6
Paraovarian cyst	1	0	1
Peritoneal cyst	1	0	1
Struma ovarii	1	0	1

Adult teratoma	10	1	11
Serous borderline tumor	2	0	2
Mucinous borderline tumor	0	1	1
Clear cell borderline tumor	1	0	1
Clear cell adenocarcinoma	1	1	2
Endometrioid adenocarcinoma	1	1	2
Adenosquamous carcinoma	0	1	1
Metastatic ovarian tumor	2	3	5
Mucinous adenocarcinoma - 2	1	1	2
Serous adenocarcinoma - 18	7	18	25

Table 2. ADNEX performance measures for all patients, pre- and postmenopausal cases depending on the Center

	All patients	95% CI	Premenopausal	95% CI	Postmenopausal	95% CI
Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland (Centre I)						
ACC (%)	79.9	74. 8 -85.3	81.2	74.6 -87.7	77.3	66.7 -86.4
SENS (%)	94.3	88.5 -98.7	90.6	77.0 -100.0	97.4	91.7 - 100.0
SPEC (%)	72.4	65.1 -79.7	78.3	70.7 -85.9	50.0	32.1 -69.8
PPV (%)	64.1	54.6 -74.2	55.8	40.7 -67.7	72.5	59.6 -83.9
NPV (%)	96.0	92.0 -99.1	96.5	91.9 - 100.0	93.3	76.9 - 100.0
AUC	0.907	0.858 -0. 948	0.886	0.789 -0.957	0.884	0.798- 0.957
Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain (Centre II)						
ACC (%)	81.3	74.0 -88.6	84.7	76.4 -93.1	76.5	64.7 -86.3
SENS (%)	97.1	89.7 - 100.0	100.0	-	95.8	85.7 - 100.0
SPEC (%)	75.3	65.2 -84.7	82.3	71.6 -91.1	59.3	41.5 -77.6
PPV (%)	60.0	47.3 -74.0	47.6	25.6 -69.4	67.6	51.4 -81.2
NPV (%)	98.5	95.1 -100.0	100.0	-	94.1	80.0 - 100.0
AUC	0.955	0.893 - 0.996	100.0	-	0.900	79.5 -98.3

The performance measures were calculated for all patients and with distinction to the menopausal status. The percentile confidence intervals were calculated with 500 times bootstrap. The "-" sign is placed when a confidence interval cannot be calculated. ACC -accuracy; SENS - sensitivity; SPEC -specificity; PPV- positive predictive value; NPV- negative predictive value; AUC - area under the receiver operating characteristic curve

Table 3. The confusion matrix for multi-class classification in Centers I and II.

	Histopathology results				
	benign	borderline	stage I	stage II-IV	metastatic
ADNEX model indications	Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland (Centre I)				
benign	97	2	0	2	0
borderline	13	4	1	11	0
stage I	13	1	0	1	1
stage II-IV	10	5	8	30	4
metastatic	1	0	0	0	0
	Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain (Centre II)				
benign	67	0	1	0	0
borderline	13	2	0	0	1
stage I	2	0	2	1	0
stage II-IV	6	2	1	19	3
metastatic	1	0	1	0	1

The columnar headings present true histopathological diagnosis. The first positions in rows refer to ADNEX indication. The multi-class accuracies were 0.642 (95% CI 0.572 - 0.708) and 0.740 (95% CI 0.653 - 0.815) for Centers I and II, respectively.

Table 4. Diagnostic measurements for multi-class recognition in Centers I and II.

	benign	borderline	stage I	stage II-IV	metastatic
Division of Gynecologic Surgery, Poznan University of Medical Sciences, Poland (Centre I)					
SENS (%)	72.4	33.3	00.0	68.2	00.0
SPEC (%)	94.3	87.0	91.8	83.1	99.5
PPV (%)	96.0	13.8	00.0	52.6	00.0
NPV (%)	64.1	95.4	95.2	90.5	97.5
Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, University of Navarra School of Medicine, Pamplona, Spain (Centre II)					
	benign	borderline	stage I	stage II-IV	metastatic
SENS (%)	75.3	50.0	40.0	95.0	20.0
SPEC (%)	97.1	88.2	97.5	88.3	98.3
PPV (%)	98.5	12.5	40.0	61.3	33.3
NPV (%)	60.0	98.1	97.5	98.9	96.7

SENS - sensitivity; SPEC -specificity; PPV- positive predictive value; NPV- negative predictive value; AUC - area under the receiver operating characteristic curve

Table 5. The results of subjective ultrasonographic assessment for all patients, pre- and postmenopausal cases depending on the Center

	All patients	95% CI	Premenopausal	95% CI	Postmenopausal	95% CI
Division of Gynecologic Surgery - Poznan University of Medical Sciences - Poland (Centre I)						
ACC (%)	89.7	85.3 - 93.6	89.9	84.8 - 94.2	89.4	80.3 - 95.5
SENS (%)	88.6	81.0 - 94.9	78.1	62.9 - 91.7	97.4	91.2-100.0
SPEC (%)	90.3	84.9 - 95.0	93.4	88.0 - 97.3	78.6	60.9 - 92.9
PPV (%)	82.7	73.8 - 90.9	78.1	62.1 - 91.4	86.0	73.3 - 95.3
NPV (%)	93.8	89.3 - 97.4	93.4	88.1 - 98.0	95.7	85.0-100.0
Department of Obstetrics and Gynecology - Clinica Universidad de Navarra - University of Navarra School of Medicine - Pamplona - Spain (Centre II)						
ACC (%)	95.1	91.1–98.4	97.2	93.1–100.0	92.2	84.3 - 98.0
SENS (%)	97.1	90.0- 100.0	100.0	-	95.8	85.7-100.0
SPEC (%)	94.4	89.0 - 98.8	96.8	91.8–100.0	88.9	76.2-100.0
PPV (%)	86.8	75.7 - 96.7	83.3	55.6–100.0	88.5	76.0-100.0
NPV (%)	98.8	96.2-100.0	100.0	-	96.0	87.5-100.0

The performance measures were calculated for all patients and with distinction to the menopausal status. The percentile confidence intervals were calculated with 500 times bootstrap. The "-" sign is placed when a confidence interval cannot be calculated. ACC -accuracy; SENS - sensitivity; SPEC -specificity; PPV- positive predictive value; NPV- negative predictive value;