The performance of selected models for predicting malignancy in ovarian tumors in relation to the degree of diagnostic uncertainty in subjective ultrasonographic assessment

Research Article

Running Title: Predictive models and diagnostic uncertainty

Sebastian Szubert MD, PhD[1,2], Dariusz Szpurek MD, PhD[3], Andrzej Wójtowicz MSc, PhD[4], Patryk Zywica MSc, PhD[4], Maciej Stukan MD, PhD[5], Stefan Sajdak MD, PhD[6], Sławomir Jabłonski MD, PhD[1], Łukasz Wicherek MD, PhD[2], Rafał Moszyński MD, PhD[6]

[1] Clinical Department of Gynaecological Oncology, The Franciszek Lukaszczyk Oncological Center, Bydgoszcz, Poland

[2] 2nd Department of Obstetrics and Gynecology, Medical Centre of Postgraduate Education, Warsaw, Poland

[3] Private Medical Practice Dariusz Szpurek, 32/4 Chwialkowskiego St., 61-553 Poznan, Poland

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

[5] Department of Gynecologic Oncology, Gdynia Oncology Center, Pomeranian Hospitals, Gdynia, Poland

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

Corresponding Author:

Sebastian Szubert, MD, PhD

Clinical Department of Gynaecological Oncology, The Franciszek Lukaszczyk Oncological Center, Bydgoszcz, Poland

2 Izabela Romanowska Street, Bydgoszcz 85-796, Poland.

Email: szuberts@o2.pl

ORCID: 0000-0003-3313-7188

Abstract

Introduction:

The study's main aim was to evaluate the relationship between the performance of predictive models for differential diagnoses of ovarian tumors and levels of diagnostic confidence in subjective ultrasonographic assessment (SA). The second aim was to identify the parameters that differentiate between malignant and benign tumors among tumors initially diagnosed as uncertain in SA.

Material and methods

The study included 250 (55%) benign ovarian masses and 201 (45%) malignant tumors. In ultrasonographic ultrasonography, the tumors were divided into six groups: certainly benign (CB), probably benign (PB), uncertain but benign (UB), uncertain but malignant (UM), probably malignant (PM) and certainly malignant (CM). The performance of the Risk of Malignancy Index (RMI), International Ovarian Tumor Analysis (IOTA) ADNEX model, and IOTA logistic regression model 2 (LR2) were analyzed in subgroups as follows: SA-certain

tumors (including CB and CM) vs. SA-probable (PB and PM) vs. SA-uncertain (UB and UM).

Results

We found a progressive decrease in the performance of all models in association with the increased uncertainty in SA. The AUC for the RMI, LR2 and ADNEX models decreased between the SA-certain and SA-uncertain groups for 20%, 28%, and 20% respectively. The presence of solid parts and a high color score were the discriminatory features between UB and UM tumors.

Conclusions

Studies are needed that focus on the subgroup of ovarian tumors that are difficult to classify in SA. In cases of uncertain tumors in SA, the presence of solid components or high color score should prompt a gynecologic oncology clinic referral.

Key words: Ovarian cancer; Ovarian tumor; Ultrasound, Subjective assessment, Predictive models

Introduction

Differential diagnosis of ovarian tumor remains a recurrent problem in gynecological practice. After diagnosis of an ovarian tumor the clinician must make the decision whether the patient requires surgical treatment, or she can be managed expectantly. Furthermore, if surgery is indicated, another issue to be resolved is whether the patient should be operated on in a specialized gynecological oncology center, or she may undergo treatment in a general gynecologic unit with a minimally invasive approach. Currently, ultrasonography with subjective assessment (SA) performed by an experienced sonographer is regarded as the most precise and specific method for the differential diagnosis of ovarian tumors [1,2]. SA is superior to other diagnostic methods such as RMI or ROMA, which also use the analysis of cancer serum biomarkers [1,3,4] Additionally, SA conducted by an expert is used when other diagnostic tests yield inconclusive results [5,6]. SA by an experienced sonographer is not only used to differentiate benign from malignant tumors. Nowadays, with more specific imaging available, recognition is easier. SA may suggest a very specific diagnosis, for example, beyond simple

differentiation, it may indicate a borderline ovarian tumor or a secondary ovarian malignancy, thus an individualized treatment approach may be applied as a result [7–10]. However, for many patients there is limited access to SA by an experienced sonographer because there is a relatively small number of gynaecological ultrasound specialists. Therefore, multiple diagnostic and predictive models, based on ultrasonography, clinical variables and cancer biomarker assessment, have been developed to better facilitate the evaluation and diagnosis of tumors. The idea behind the development of predictive models for a differential diagnosis of ovarian tumors was to enable inexperienced sonographers to undertake diagnoses [11,12]. In that context, a physician who is less experienced in gynecologic ultrasound, has at their disposal another diagnostic tool for differentiating malignant from benign ovarian tumors. Therefore, it could be said that the relative experience of the sonographer determines whether there is a need to apply a predictive model. However, every sonographer has at least some experience in differentiating ovarian tumors in SA. Further, it is true that multiple benign ovarian tumors (for instance, most endometriosis cysts and dermoids) and evident malignancies (i.e., advanced ovarian cancers) are easy to recognise, even by beginners. In such situations predictive models are redundant.

Predictive models for the differential diagnosis of ovarian tumors require prospective validation before clinical application. Most studies report that internal validation is performed at the time of the original reports. In general, the studies provide detailed characterizations of the tumors (the ultrasonographic structure, and histopathological type, etc.); however, data is sparse about the level of diagnostic confidence in relationship to the SA of the tumor [12–15]. This is of clinical significance, because from a practical point of view, the predictive models should prove to have been effective when using SA by a non-expert is unequivocal. We hypothesize that as diagnostic certainty decreases in SA, and therefore, as uncertainty

increases, the accuracy of the other diagnostic tests also decreases. Thus, the main aim of our study was to evaluate the diagnostic performance of selected diagnostic models in relation to the degree of uncertainty in SA.

Materials and method

Informed consent was obtained from all individual participants included in the study. The study was approved by the Poznan University of Medical Science Ethics Committee (884/17). We retrospectively evaluated data collected from the ultrasonographic database of ovarian tumors in patients who had been referred to our clinics. In matter of the material in the study that was sourced from the Division of Gynecologic Surgery, of the Poznan University of Medical Sciences, Poland, the data had been obtained from patients treated for ovarian tumors between December 2010 and April 2018. The study included 368 consecutive women who had an ultrasonographic examination due to an ovarian tumor that was performed by either S.Sz or R.M. The study group included women were referred to S.Sz or R.M. for an ultrasonography consultation by a less-experienced physician; and others who were evaluated by S.Sz or R.M. because the women were admitted to the hospital on one of these physician's routine duty days. Ultrasonography was performed according to the IOTA criteria for describing the sonographic morphology of ovarian tumors [16]. Only patients with CA125 data available were enrolled. There were no specific exclusion criteria, and the only inclusion criterion was the patient's need for surgery due to an ovarian tumor.

Ultrasonography was performed one to three days before surgery. The tumors were evaluated using Aloka Alpha 10 with a 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 cases of large tumors. In cases of bilateral ovarian tumors, the data of the tumor with the more

complex morphology were collected. If the tumors had similar morphologies, the data of the largest one was selected. The tumors were assessed by either R.M. or S.Sz. R.M. has over 16 years' experience in gynecological ultrasonography, having performed approximately 800 examinations per year. S.Sz. has 12 years' experience in gynecological ultrasonography and in the past two years performed 300 examinations each year, and prior to that, 1000 examinations per year. Both R.M. and S.Sz. conduct clinical studies in the field of gynecological ultrasonography and teach in numerous courses and give lectures on the field of ultrasound examinations. However, despite their experience, gynecological ultrasonography is not the main field of expertise of either S.Sz or R.M., thus, applying the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) criteria, these sonographers classify themselves as level 2 examiners.

We also included data collected from June 2016 to September 2017 at the Department of Gynecologic Oncology, Gdynia Oncology Center, of the Pomeranian Hospitals, Gdynia, Poland. The study included 83 patients with ovarian tumors who had undergone consecutive preoperative ultrasonographic examination performed by M.S. All examinations were performed 1 to 3 days before surgery using the standards and terminology proposed by the IOTA group [16]. Similarly, CA125 serum levels were evaluated 1 to 3 days prior to surgery. The patients underwent transvaginal or transrectal ultrasound using a Philips HD15 Ultrasound System with Philips C8-4v Endovaginal Probe, 4-8 MHz and Philips V6-2 broadband convex transducer, 6-2 MHz (Philips Healthcare, Koninklijke, The Netherlands). M.S. has over 20 years of experience in gynecological ultrasonography. He is the author of numerous studies concerning differential diagnosis of ovarian tumors. M.S. is a teacher of gynecological ultrasonography and he is regarded as an expert in this field. However, his

main field of expertise is gynecologic surgery; thus M.S. classifies himself as level 2 ultrasonography practitioner according to the EFSUMB criteria.

Following each ultrasound examination, the examiners indicated their subjective impression about the tumor's character, and using the IOTA rules, classified the masses as: certainly benign (CB), probably benign (PB), uncertain but benign (UB), uncertain but malignant (UM), probably malignant (PM) and certainly malignant (CM) [1718]. Our study's analysis was performed between pairs of certain (SA-certain; including CB+CM tumors), probable (SA-probable; including PB+PM tumors) and uncertain (SA-uncertain; including UB+UM) tumors because we believe the corresponding groups are similar to each other with regard to the degree of diagnostic confidence. Each SA examination was a blind test, as the examiners were not given access to the predictive model results.

All tumors were surgically removed. the reference standard was the final histopathological diagnosis obtained for all tumors using the WHO classification [19]. Borderline tumors were classified as malignant tumors. Data collected in the ultrasonographic database was used to assess the following predictive models according to the methodologies described in the source literature: risk of malignancy index (RMI) [19], logistic regression model 2 [20], and the Assessment of Different Neoplasias in the adneXa (ADNEX) [21] developed by the International Ovarian Tumor Analysis (IOTA). The cut-off for RMI was set as 200 points. In the case of the ADNEX model and LR2, a greater than 10% risk of a malignant tumor was considered as an indication of malignancy.

The test results were evaluated using the diagnostic odds ratio (DOR) and the area under the Receiver Operating Characteristic (ROC) curve (AUC) [20]. The sensitivity (SENS), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV), and the accuracy of all tests were also calculated.

Mathematical and statistical analyses were based on software R version 3.5.1 (2018-07-02) with libraries pROC v. 1.12.1. For categorical variables, independence between groups was studied using the Fisher exact test. The DeLong et. al., method was used for the comparison of AUC between subgroups [21].

The study was conducted in adherence with the 2015 guidelines of the Standards for Reporting of Diagnostic Accuracy Studies (STARD). The study received no funding.

Results

The study group included 250 benign ovarian masses (55%) and 201 (45%) malignant tumors. There were 22 (5%) borderline, 44 (10%) stage one and 126 (%) stage II-IV ovarian malignancies, and 9 (2%) secondary ovarian malignancies. Two-hundred seventy women were premenopausal (60%), while 181 (40%) were postmenopausal (postmenopausal being defined as 1 year after the last period and with no other endocrine disorders; or older than 50 years' old if they had undergone hysterectomy). Data on each patient's age, CA125 levels and tumor ultrasonographic morphology according to the type of tumor are shown in Table 1.

By the end of the study, the group included 72 (16%) certainly benign (CB), 137 (30%) probably benign (PB), 34 (8%) uncertain but benign (UB), 52 (12%) uncertain but malignant (UM), 74 (16%) probably malignant (PB) and 82 (18%) certainly malignant (CM) ovarian tumors.

The results of histopathological examinations are shown in Table 2.

The performance of the diagnostic models and the SA in groups of tumors we analyzed is presented in Table 3.

In all the models we studied, we observed lower accuracy, sensitivity, specificity, positive and negative predictive values, and DORs in the group of SA-uncertain tumors compared with the results for the SA-certain and SA-probable group.

We found significantly higher AUCs for LR2 in the group of SA-certain tumors than in both the SA-probable (P = 0.001) and SA-uncertain (P = 0.034) groups of tumors. However, there were no differences in the AUCs when we compared the LR2 model with the SA-probable and SA-uncertain groups of tumors (P = 0.549). At the same time, we found significantly higher AUCs for the ADNEX model in the SA-certain tumors group when compared with the SA-probable (P = 0.012) and SA-uncertain groups of tumors (P = 0.034). The difference in the AUCs for the ADNEX model comparing the SA-probable and SA-uncertain groups of tumors was insignificant (P = 0.635). We found no significant differences in the AUCs for RMI when its performance was compared between the groups of tumors we studied. The P-values for the comparisons of the AUCs for RMI between the groups studied were as follows: P = 0.122 for SA-certain vs SA-probable tumors; P=0.108 for SA-certain vs SA-uncertain tumors, and P = 0.146 for SA-probable vs SA-uncertain tumors. The AUC for RMI, LR2 and ADNEX decreased between the SA-certain and SA-uncertain tumors by 20%, 28% and 20% respectively. While, the corresponding decreases of the AUC between the SA-probable and SA-uncertain tumors was 11%, 6% and 11% respectively.

When all six groups of tumors were taken into consideration, we found statistically significant differences in the patients' ages, CA-125 levels and the ultrasonographic features between the levels of diagnostic confidence pertaining to the groups of tumors classified in SA. Detailed results are presented in the supplementary Table 1. When we subsequently focused on differentiating between UB and UM tumors, we found solid parts more frequently in UM than in UB tumors (P < 0.001). Additionally, UM tumors had a significantly higher median color

score when compared with UB tumors (4, range 2-4 vs 2, range 1-3; P = 0.008). We found no significant difference between UB and UM tumors in the other ultrasonographic features that were analyzed. Furthermore, there were no differences between the groups in terms of the patients' ages, the CA-125 levels, or menopausal status. The results of the comparisons between UB and UM tumors are summarized in Table 4.

Discussion

Predictive models for the differential diagnosis of ovarian tumors were developed mainly to facilitate diagnosis when experienced sonographic assessment is unavailable. Thus, in practice, the diagnostic models should improve decision making. However, in our study we observed a progressive decrease in the performance of predictive models for the differential diagnosis of ovarian tumors, along with an increased uncertainty with subjective ultrasonographic assessment. The reduced quality of performance was observed in all of predictive models we studied (RMI, LR2 and ADNEX) and presented as declines in the AUCs, DORs and the sensitivity and specificity of the diagnostic tool. The poor performance of the models was observed in both the uncertain tumors group, as well as in the group of probably benign and probably malignant tumors. In the cases of tumors where the observer had no doubt about the character of the tumor, we found that all of the tumors were classified correctly by SA and all of the predictive models studied performed at an excellent level. On the other hand, when the diagnosis was difficult in SA, the performances of the predictive models was also found to be lower. The results of our study point out important issues about other studies on predictive models for ovarian tumors and the clinical utility of the models. Firstly, we consider, when the predictive models are assessed, it seems reasonable to provide the data about the level of diagnostic confidence in SA for the tumors included. In general,

other studies on the efficacy of prognostic models provide detailed characteristics of sonographic features and the clinical data on the women in the studies [22–24]. Data about relative confidence levels of the subjective assessment would provide information about the clinical difficulties encountered in the diagnosis of the tumors included in the studies, thereby providing essential information about the conditions under which the predictive model was validated. Secondly, it would be worthwhile evaluating the true clinical utility of predictive models for ovarian tumors, because our study shows their performance is weaker in those situations where they are needed the most.

In recent years, numerous predictive models and tests have been developed for the differential diagnosis of adnexal masses. From a practical point of view, it would be of clinical interest to distinguish those models which are useful for differential diagnosis specifically for the group of adnexal tumors which are difficult to assess. In a study by Valentin et al., the authors of the large multicenter study reported that 7% of adnexal tumors could not be classified by an experienced sonographer in SA, as either benign or malignant [17]. In our study group, 20% of the tumors studied constituted the subgroup of tumors that were difficult to diagnose in SA (UB and UM). This incidence of uncertain tumors, a higher percentage than in the cited study, may have been a result of the character of the tumors we studied; given that the study group was of ovarian tumors, most of which were malignant, and which had been referred to the reference center for gynecological surgery for surgery. Furthermore, significant proportions of the tumors we studied had been sent to us by other physicians for expert consultation. Finally, we presume that our experience is at a lower level than the highly experienced experts in the IOTA group. The diagnostics of difficult tumors in SA remains a persistent problem in gynecology. In the study by Valentin et al., cited above, the authors developed a logistic regression model to differentiate the unclassifiable adnexal tumors [17]. However, the logistic

302 regression model, as well as the RMI and CA125 levels assessment, failed to differentiate
303 benign from malignant tumors in their subgroup of unclassifiable adnexal tumors [17]. In
304 previous study that we published, we found the evaluation of HE4 levels as a useless
305 additional test for evaluating uncertain adnexal tumors in SA [25]. In our present study we have
306 found that all the predictive models we studied had similar DORs and AUCs within the
307 uncertain tumors group. However, due to the limited number of cases in our subgroup of
308 uncertain tumors, we did not set out to compare the models, but to show the rule of the
309 decreased performance of the predictive model in conjunction with an increased uncertainty
310 in SA.
311 In the study by Valentin et al., borderline tumors, fibromas, and serous and mucinous
312 cystadenoma/cystadenofibroma were the most common among the unclassifiable masses.
313 Similarly, those types of tumor were significantly more commonly classified incorrectly as
314 benign or malignant, when compared with the other tumors in their study [17]. The authors
315 compared the ultrasonographic characteristics of the unclassifiable with the classifiable
316 adnexal masses. The former group of tumors were found to be larger, more often had a
317 unilocular-solid, multilocular or multilocular-solid appearance, and more often had an
318 irregular wall and papillary projections when compared with the latter tumors. The
319 unclassifiable tumors also had fewer papillary projections, smaller solid components, and
320 more commonly presented with moderate vascularization (Color score 3). In our study we
321 preferred to compare the ultrasonographic features of the tumors divided into six sub-
322 categories according to the levels of diagnostic confidence in SA. We found that the group of
323 tumors categorized as difficult to classify in SA shared intermediate features with those
324 tumors classified at the two boundaries of diagnostic confidence. That indicates, that the
325 group of difficult to classify tumors in SA include the features of both malignant and benign

tumors, therefore making them difficult to classify both in SA and with predictive models. Next we focused on differentiating between the UB and UM tumors. Here we found, that the presence of solid components and high color scores were the discriminatory features between the UB and UM tumors. However, more than half of the UB tumors were also found to have solid tumor elements. When considering the color scores, one-third of the UB tumors (35%) were moderately (score 3) or highly (score 4) vascularized. In the previously cited study by Valentin et al., the only variable used in their multivariate regression model to calculate the risk of malignancy among unclassifiable ovarian tumors was the diameter of the largest solid components [17]. However, the logistic regression model they developed performed weakly when discriminating between malignant and benign ovarian tumors in the subgroup of unclassified tumors [17]. The management of indeterminate ovarian masses remains a persistent problem in gynecology. The First International Consensus Report on Adnexal Masses includes a "next steps" proposition when the diagnosis of an indeterminate ovarian tumor is established. However, in the end, referral to a gynecologic oncologist for surgical evaluation remains a reasonable option [26].

To the best of our knowledge, this is the first study reporting the relationship between the degrees of uncertainty in SA with the performance levels of predictive models. The advantage of our study is that it was conducted in two centers, included a significant number of patient cases, and involved comprehensive ultrasonographic assessment of the tumors. Additionally, the performance of the predictive models was analyzed using their sensitivity, specificity, negative and positive predictive values as well as the AU-ROCs and DORs. However, the study does have some limitations. The main limitations of this study include its retrospective character. Additionally, the proportion of malignant to benign ovarian tumors reported in our study is a reflection of the proportion that is characteristic of gynecologic oncology clinics,

and does not therefore reflect the actual incidence ratios of malignant and benign ovarian tumors. Furthermore, the degree of diagnostic confidence is very subjective and is strictly related to the relative experience of examiners. We did not perform an analysis of the various cut-offs and the calibration of analyzed models, because the aim of our study was not to evaluate their performance, but to show the relationship between the performances of the various models and the diagnostic confidence in SA.

Conclusions

Implications for research:

We propose that, because of the significantly weaker diagnostic performance of the diagnostic models with the tumors in the difficult to classify as benign or malignant group in SA, future clinical studies should give additional attention to this subgroup of ovarian tumors. Furthermore, when new predictive models are developed, or, the validation of existing models is tested, it would be reasonable to include, along with the characteristics of the ovarian tumors, data concerning the levels of diagnostic confidence in SA.

Implications for practice:

In cases of uncertain tumors in SA, the presence of solid components or abundant tumor vasculature (high color score) should prompt referrals to a gynecologic oncology clinic.

Acknowledgement

We would like also to thank Robert Garrett for his assistance with the manuscript.

References

1. Van Gorp T, Veldman J, Van Calster B, et al. Subjective assessment by ultrasound is superior to the risk of malignancy index (RMI) or the risk of ovarian malignancy algorithm (ROMA) in discriminating benign from malignant adnexal masses. *Eur J Cancer*. 2012;48(11):1649-1656. doi:10.1016/j.ejca.2011.12.003
2. Valentin L, Jurkovic D, Van Calster B, et al. Adding a single CA 125 measurement to ultrasound imaging performed by an experienced examiner does not improve preoperative discrimination between benign and malignant adnexal masses. *Ultrasound Obstet Gynecol*. 2009;34(3):345-354. doi:10.1002/uog.6415
3. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol*. 1990;97(10):922-929. http://www.ncbi.nlm.nih.gov/pubmed/2223684. Accessed July 1, 2019.
4. Moore RG, Miller MC, Disilvestro P, et al. Evaluation of the diagnostic accuracy of the risk of ovarian malignancy algorithm in women with a pelvic mass. *Obstet Gynecol*. 2011;118(2 Pt 1):280-288. doi:10.1097/AOG.0b013e318224fce2
5. Timmerman D, Ameye L, Fischerova D, et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. *BMJ*. 2010;341:c6839. doi:10.1136/bmj.c6839
6. Alcázar JL, Pascual MÁ, Olartecoechea B, et al. IOTA simple rules for discriminating between benign and malignant adnexal masses: prospective external validation. *Ultrasound Obstet Gynecol*. 2013;42(4):n/a - n/a. doi:10.1002/uog.12485
7. Yazbek J, Ameye L, Testa AC, et al. Confidence of expert ultrasound operators in

making a diagnosis of adnexal tumor: effect on diagnostic accuracy and interobserver agreement. *Ultrasound Obstet Gynecol*. 2010;35(1):89-93. doi:10.1002/uog.7335

8. Valentin L. Use of morphology to characterize and manage common adnexal masses. *Best Pract Res Clin Obstet Gynaecol*. 2004;18(1):71-89. doi:10.1016/j.bpobgyn.2003.10.002

9. Pascual A, Guerriero S, Rams N, et al. Clinical and ultrasound features of benign, borderline, and malignant invasive mucinous ovarian tumors. *Eur J Gynaecol Oncol*. 2017;38(3):382-386. http://www.ncbi.nlm.nih.gov/pubmed/29693878. Accessed May 28, 2019.

10. Fischerova D, Zikan M, Dundr P, Cibula D. Diagnosis, treatment, and follow-up of borderline ovarian tumors. *Oncologist*. 2012;17(12):1515-1533. doi:10.1634/theoncologist.2012-0139

11. Sayasneh A, Kaijser J, Preisler J, et al. Accuracy of ultrasonography performed by examiners with varied training and experience in predicting specific pathology of adnexal masses. *Ultrasound Obstet Gynecol*. 2015;45(5):605-612. doi:10.1002/uog.14675

12. Wynants L, Timmerman D, Verbakel JY, et al. Clinical Utility of Risk Models to Refer Patients with Adnexal Masses to Specialized Oncology Care: Multicenter External Validation Using Decision Curve Analysis. *Clin Cancer Res*. 2017;23(17):5082-5090. doi:10.1158/1078-0432.CCR-16-3248

13. Van Holsbeke C, Van Calster B, Testa AC, et al. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: Results from the international ovarian tumor analysis study. *Clin Cancer Res*. 2009;15(2):684-691. doi:10.1158/1078-0432.CCR-08-0113

14. Sayasneh A, Wynants L, Preisler J, et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer*. 2013;108(12):2448-2454. doi:10.1038/bjc.2013.224

15. Stukan M, Badocha M, Ratajczak K. Development and validation of a model that includes two ultrasound parameters and the plasma D-dimer level for predicting malignancy in adnexal masses: an observational study. *BMC Cancer*. 2019;19(1):564. doi:10.1186/s12885-019-5629-x

16. 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(5):500-505. doi:10.1046/j.1469-0705.2000.00287.x

17. Valentin L, Ameye L, Savelli L, et al. Adnexal masses difficult to classify as benign or malignant using subjective assessment of gray-scale and Doppler ultrasound findings: logistic regression models do not help. *Ultrasound Obstet Gynecol*. 2011;38(4):456-465. doi:10.1002/uog.9030

18. Testa A, Kaijser J, Wynants L, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. *Br J Cancer*. 2014;111(4):680-688. doi:10.1038/bjc.2014.333

19. Meinhold-Heerlein I, Fotopoulou C, Harter P, et al. The new WHO classification of ovarian, fallopian tube, and primary peritoneal cancer and its clinical implications. *Arch Gynecol Obstet*. 2016;293(4):695-700. doi:10.1007/s00404-016-4035-8

20. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77.

doi:10.1186/1471-2105-12-77

21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845. http://www.ncbi.nlm.nih.gov/pubmed/3203132. Accessed October 23, 2019.

22. Alcázar JL, Mercé LT, Laparte C, Jurado M, López-García G. A new scoring system to differentiate benign from malignant adnexal masses. *Am J Obstet Gynecol*. 2003;188(3):685-692. doi:10.1067/mob.2003.176

23. Moszynski R, Zywica P, Wojtowicz 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(12):892-899.

24. Szubert S, Wojtowicz A, Moszynski R, et al. External validation of the IOTA ADNEX model performed by two independent gynecologic centers. *Gynecol Oncol*. 2016;142(3):490-495. doi:10.1016/j.ygyno.2016.06.020

25. Moszynski R, Szubert S, Szpurek D, Michalak S, Krygowska J, Sajdak S. Usefulness of the HE4 biomarker as a second-line test in the assessment of suspicious ovarian tumors. *Arch Gynecol Obstet*. 2013;288(6):1377-1383. doi:10.1007/s00404-013-2901-1

26. Glanc P, Benacerraf B, Bourne T, et al. First International Consensus Report on Adnexal Masses: Management Recommendations. *J Ultrasound Med*. 2017;36(5):849-863. doi:10.1002/jum.14197

Table 1. Clinical and ultrasound ovarian tumor characteristics according to the reference index of the ovarian tumor

	Benign 250 (55%)	Borderline 22 (5%)	stage I 44 (10%)	stage II-IV 126 (28%)	Metastatic 9 (2%)
	Median (inter-quartile range)				
Age	42 (31-53)	52.5 (32-64)	51.5 (44-63)	58.5 (51-65)	53 (46-53)
CA-125	27.5 (15-58)	36.315 (18-115)	226.865 (68-1024)	506 (167-1476)	139.5 (84-542)
Lesion maximal diameter	65 (51-100)	90.5 (55-170)	117 (94-152)	100 (51-134)	105 (85-180)
Solid part maximal diameter	0 (0-20)	19.5 (12-51)	50 (24-54)	50 (32-74)	50 (45-57)
	Number (%)				
Presence of solid parts	93 (37%)	18 (82%)	39 (89%)	123 (98%)	8 (89%)
More than 10 locules	23 (9%)	7 (32%)	10 (23%)	21 (17%)	1 (11%)
Acoustic shadows	22 (9%)	1 (5%)	0 (0%)	6 (5%)	0 (0%)
Ascites	15 (6%)	2 (9%)	10 (23%)	73 (58%)	3 (33%)
	Number of papillary projections N (%)				
0	147 (59%)	6 (27%)	18 (41%)	66 (52%)	5 (56%)
1	26 (10%)	2 (9%)	3 (7%)	12 (10%)	1 (11%)
2	27 (11%)	2 (9%)	4 (9%)	6 (5%)	1 (11%)
3	24 (10%)	4 (18%)	5 (11%)	9 (7%)	1 (11%)
more than 3	26 (10%)	8 (36%)	14 (32%)	33 (26%)	1 (11%)

Table 2. Distribution of histopathological findings among tumors included in the study

	Premenopausal N (%)	Postmenopausal N (%)	All N (%)
adenofibroma	4 (0.9%)	4 (0.9%)	8 (1.8%)
adult teratoma	30 (6.7%)	5 (1.1%)	35 (7.8%)
Brenner tumor	12 (2.7%)	10 (2.2%)	22 (4.9%)
corpus luteum cyst	0 (0.0%)	2 (0.4%)	2 (0.4%)
endometrioid cyst	4 (0.9%)	4 (0.9%)	8 (1.8%)
granulosa cell tumor	2 (0.4%)	2 (0.4%)	4 (0.9%)
hemorrhagic cyst	11 (2.4%)	11 (2.4%)	22(4.9%)
mucinous cystadenoma	78 (17.3%)	4 (0.9%)	82 (18.2%)
peduculated leiomyoma	4 (0.9%)	0 (0.0%)	4 (0.9%)
serous cystadenoma	6 (1.3%)	0 (0.0%)	6 (1.3%)
simple cyst	5 (1.1%)	4 (0.9%)	9 (2.0%)
theca cell tumor	3 (0.7%)	3 (0.7%)	6 (1.3%)
tubo-ovarian abscess	19 (4.2%)	11 (2.4%)	30 (6.7%)
borderline tumor	3 (0.7%)	1 (0.2%)	4 (0.9%)
clear cell adenocarcinoma	37 (8.2%)	70 (15.5%)	107 (23.7%)
endometrioid adenocarcinoma	18 (4.0%)	24 (5.3%)	42 (9.3%)
mucinous adenocarcinoma	14 (3.1%)	5 (1.1%)	19 (4.2%)
serous adenocarcinoma	2 (0.4%)	5 (1.1%)	7 (1.6%)
metastatic ovarian tumor	10 (2.2%)	1 (0.2%)	11 (2.4%)
undifferentiated carcinoma	8 (1.8%)	15 (3.3%)	23 (5.1%)
Total	270 (59.9%)	181 (40.1%)	451 (100.0%)

Table 3. The performance of diagnostic models and subjective assessment (SA) within the subgroups of ovarian tumors analyzed

	model	ACC [95% CI]	SEN [95% CI]	SPEC [95% CI]	PPV [95% CI]	NPV [95% CI]	DOR [range]	AUC [95% CI]
SA- certain tumors	RMI	0.925 [0.879 - 0.969]	0.862 [0.767 - 0.938]	0.986 [0.952 - 1.000]	0.982 [0.938 - 1.000]	0.883 [0.803 - 0.951]	423.111 [88.566 -1026.682]	0.989 [0.977- 0.989]
	LR2	0.927 [0.874 - 0.972]	1 [1-1]	0.886 [0.803 - 0.955]	0.83 [0.711 - 0.930]	1 [1-1]	NA	0.981 [0.945- 0.981]
	Adnex	0.87 [0.802 - 0.925]	1 [1-1]	0.775 [0.667 - 0.864]	0.765 [0.652 - 0.863]	1 [1-1]	NA	1 [1-1]
	SA	1 [1-1]	1 [1-1]	1 [1-1]	1 [1-1]	1 [1-1]	NA	1 [1-1]
SA- probable tumors	RMI	0.833 [0.786 - 0.880]	0.761 [0.657 - 0.861]	0.869 [0.817 - 0.923]	0.739 [0.639 - 0.838]	0.881 [0.828 - 0.936]	21.073 [11.369 -47.633]	0.888 [0.835 - 0.888]
	LR2	0.759 [0.693 - 0.821]	0.922 [0.849 - 0.984]	0.66 [0.567 - 0.755]	0.621 [0.522 - 0.725]	0.933 [0.870 - 0.986]	22.944 [9.601 - 95.200]	0.743 [0.675 - 0.743]
	Adnex	0.59 [0.519 - 0.663]	0.969 [0.915 - 1.000]	0.414 [0.331 - 0.504]	0.434 [0.349 - 0.517]	0.967 [0.909 - 1.000]	22.28 [6.959 - 59.468]	0.89 [0.842-0.89]
	SA	0.891 [0.848 - 0.929]	0.87 [0.783 - 0.938]	0.901 [0.855 - 0.946]	0.811 [0.719 - 0.897]	0.934 [0.887 - 0.971]	60.952 [28.057 - 176.387]	0.885 [0.839 - 0.885]
SA- uncertain tumors	RMI	0.741 [0.651 - 0.835]	0.694 [0.549 - 0.818]	0.806 [0.676, 0.930]	0.829 [0.721 - 0.935]	0.659 [0.520 - 0.795]	9.39 [3.916 - 34.627]	0.796 [0.697 - 0.796]
	LR2	0.7 [0.600, 0.800]	0.917 [0.830 - 0.981]	0.375 [0.212 - 0.564]	0.688 [0.574 - 0.797]	0.75 [0.500, 0.947]	6.6 [2.000 - 33.726]	0.703 [0.584 - 0.703]
	Adnex	0.619 [0.506 - 0.718]	0.98 [0.932 - 1.000]	0.088 [0.000 - 0.200]	0.612 [0.500, 0.714]	0.75 [0.000 - 1.000]	4.742 [0.000 - 10.911]	0.796 [0.699 - 0.796]
	SA	0.721 [0.628 - 0.814]	0.78 [0.660 - 0.894]	0.639 [0.486 - 0.800]	0.75 [0.622 - 0.863]	0.676 [0.515 - 0.844]	6.273 [2.543 - 21.612]	0.709 [0.611 - 0.709]

ACC – accuracy; ADNEX - Assessment of Different Neoplasiasin the adneXa (ADNEX) developed by the IOTA group; AUC - area under the receiver operating characteristic (ROC) curve (AUC); DOR – diagnostic odds ratio; LR2 - logistic regression model 2 by the International Ovarian Tumor Analysis (IOTA) group; NA – not available; NPV – negative predictive value; PPV – positive predictive value; RMI – risk of malignancy index; SA – subjective assessment by an ultrasonographer; SA-certain – refers to ovarian tumors assessed as certainly malignant or certainly benign in SA; SA-probable – refers to ovarian tumors assessed as probably malignant or probably benign in SA; SA-uncertain – refers to ovarian tumors assessed as uncertain in SA, and finally classified as uncertain but malignant, or uncertain but benign; SEN – sensitivity, SPEC – specificity; 95% CI - 95% confidence interval.

Table 4. The comparison of ultrasonographic features, CA-125 levels and patient characteristics between the ovarian tumors assessed as uncertain but malignant (UM) and as uncertain but benign (UB) in subjective assessment.

	uncertain malignant (UM) N = 52	uncertain benign (UB) N = 34	
	Median (inter-quartile range)		p-value
Age	54.5 (46-60)	42 (34-56)	0.212
CA-125	146.25 (34-584)	35.46 (18-65)	0.554
Lesion max diameter	106.5 (69-150)	107.5 (61-189)	0.379
Solid part max diameter	45 (22-50)	12 (0-34)	0.067
Presence of solid parts	48 (92%)	19 (56%)	< 0.001
More than 10 locules	10 (19%)	6 (18%)	1
Acoustic shadows	1 (2%)	2 (6%)	0.559
Ascites	15 (29%)	4 (12%)	0.07
Color score	3 (2-4)	2 (1-3)	0.008
number of papillary projections	Number (%)		p-value
0	19 (22%)	14 (16%)	0.848
1	5 (6%)	5 (6%)	
2	9 (10%)	6 (7%)	
3	6 (7%)	2 (3%)	
more than 3	13 (15%)	7 (8%)	
Tumor classification	Number (%)		p-value
unilocular	1 (1%)	2 (2%)	0.061

unilocular solid	5 (6%)	9 (10%)	
Multilocular	8 (9%)	5 (6%)	
Multilocular solid	26 (30%)	16 (19%)	
solid	12 (14%)	2 (3%)	
Color score	Number (%)		p-value
1	8 (9%)	14 (16%)	0.008
2	17 (20%)	8 (9%)	
3	4 (5%)	6 (7%)	
4	23 (27%)	6 (7%)	

601

Supplementary Table 1. Patient's age, CA125 levels and ultrasonographic features of the tumors from subgroups divided according to the subjective assessment.

	Certain malignant (CM) N = 82 (18%)	Probably malignant (PM) N= 74 (16%)	Uncertain but malignant (UM) N = 52 (12%)	Uncertain but benign (UB) N = 34 (8%)	Probably benign (PB) N = 137 (30%)	Certain benign (CB) N = 72 (16%)	p-value
	Median (inter-quartile range)						
Age	59 (53-68)	53 (48-64)	54.5 (46-60)	42 (34-56)	41 (29-51)	41 (33-50)	< 0.001
CA-125	486.5 (162-1476)	228.1 (59-976)	146.25 (34-584)	35.46 (18-65)	28 (14-57)	23.7 (14-56)	0.128
Lesion max diameter	90 (46-121)	126.5 (100-179)	106.5 (69-150)	107.5 (61-189)	67 (52-94)	52 (41-68)	< 0.001
Solid part max diameter	56.5 (36-81)	50 (20-64)	45 (22-50)	12 (0-34)	0 (0-20)	0 (0-0)	< 0.001
Presence of solid parts	82 (100%)	69 (93%)	48 (92%)	19 (56%)	48 (35%)	15 (21%)	< 0.001
Color score	3 (2-3)	4 (3-4)	3 (2-4)	2 (1-3)	1 (1-2)	1 (1-1)	< 0.001
	Number (%)						
More than 10 locules	7 (9%)	26 (35%)	10 (19%)	6 (18%)	10 (7%)	3 (4%)	< 0.001
Acoustic shadows	4 (5%)	1 (1%)	1 (2%)	2 (6%)	16 (12%)	5 (7%)	0.043
Ascites	48 (59%)	29 (39%)	15 (29%)	4 (12%)	7 (5%)	0 (0%)	< 0.001
	Color score number (%)						
1	14 (3%)	5 (1%)	8 (2%)	14 (3%)	87 (19%)	64 (14%)	P <0.001
2	19 (4%)	9 (2%)	17 (4%)	8 (2%)	33 (7%)	4 (1%)	
3	29 (6%)	14 (3%)	4 (1%)	6 (1%)	11 (2%)	1 (0%)	
4	18 (4%)	41 (9%)	23 (5%)	6 (1%)	2 (0%)	0 (0%)	
	number of papillary projections number (%)						
0	50 (11%)	27 (6%)	19 (4%)	14 (3%)	68 (15%)	64 (14%)	P < 0.001)
1	9 (2%)	7 (2%)	5 (1%)	5 (1%)	12 (3%)	6 (1%)	
2	2 (0%)	4 (1%)	9 (2%)	6 (1%)	18 (4%)	1 (0%)	

3	2 (0%)	12 (3%)	6 (1%)	2 (0%)	20 (4%)	1 (0%)	
more than 3	19 (4%)	24 (5%)	13 (3%)	7 (2%)	19 (4%)	0 (0%)	
			Type of the tumor number (%)				
multilocular	8 (2%)	6 (1%)	8 (2%)	5 (1%)	22 (5%)	7 (2%)	P < 0.001)
multilocular solid	31 (7%)	48 (11%)	26 (6%)	16 (4%)	20 (4%)	3 (1%)	
notclassifiable	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0%)	0 (0%)	
solid	31 (7%)	15 (3%)	12 (3%)	2 (0%)	11 (2%)	3 (1%)	
unilocular	0 (0%)	1 (0%)	1 (0%)	2 (0%)	38 (8%)	54 (12%)	
unilocular solid	0 (0%)	2 (0%)	5 (1%)	9 (2%)	42 (9%)	3 (1%)	

4