PRACE ORYGINALNE
*ginekologia*

# Menopausal status strongly influences the utility of predictive models in differential diagnosis of ovarian tumors: An external validation of selected diagnostic tools

## Status menopauzalny – główny czynnik determinujący dokładność prognostyczną modeli diagnostyki różnicowej guzów przydatków

Rafal Moszynski[1], Patryk Zywica[2], Andrzej Wojtowicz[2], Sebastian Szubert[1], Stefan Sajdak[1], Anna Stachowiak[2], Krzysztof Dyczkowski[2], Maciej Wygralak[2], Dariusz Szpurek[1]

[1] Division of Gynecological Surgery, Poznan University of Medical Sciences, Poznan, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznan, Poland

### Abstract

***Objectives:*** *The aim of this study was to externally validate the diagnostic performance of the International Ovarian Tumor Analysis logistic regression models (LR1 and LR2, 2005) and other popular prognostic models including the Timmerman logistic regression model (1999), the Alcazar model (2003), the risk of malignancy index (RMI, 1990), and the risk of malignancy algorithm (ROMA, 2009). We compared these models to subjective ultrasonographic assessment performed by an experienced ultrasonography specialist, and with our previously developed scales: the sonomorphologic index and the vascularization index. Furthermore, we evaluated diagnostic tests with regard to the menopausal status of patients.*

***Materials and methods:*** *This study included 268 patients with adnexal masses; 167 patients with benign ovarian tumors and 101 patients with malignant ovarian tumors were enrolled. All tumors were evaluated by using transvaginal ultrasonography according to the diagnostic criteria of the analyzed models.*

*Materials and methods: This study included 268 patients with adnexal masses; 167 patients with benign ovarian tumors and 101 patients with malignant ovarian tumors were enrolled. All tumors were evaluated by using transvaginal ultrasonography according to the diagnostic criteria of the analyzed models.*

***Results:*** *The subjective ultrasonographic assessment and all of the studied predictive models achieved similar diagnostic performance in the whole study population. However, significant differences were observed when pre- and postmenopausal patients were analyzed separately. In the subgroup of premenopausal patients, the highest area under the curve (AUC) was achieved by subjective ultrasonographic assessment (0.931), the Alcazar model (0.912), and LR1 (0.909). Alternatively, in the group of postmenopausal patients, the highest AUC was noted for the Timmerman model (0.973), ROMA (0.951), and RMI (0.938).*

**Corresponding Author:**
Sebastian Szubert
Division of Gynecological Surgery, 33. Polna St.; 60-535 Poznan, Poland,
Phone: +48 61 8419490; Fax: +48 61 8419418
E-mail: szuberts@o2.pl

Otrzymano: **25.06.2014**
Zaakceptowano do druku: **25.07.2014**


© Polskie Towarzystwo Ginekologiczne

***Conclusions:*** *Menopausal status is a key factor that affects the utility of prognostic models for differential diagnosis of ovarian tumors. Diagnostic models of ovarian tumors are reasonable tools for predicting tumor malignancy.*

Key words: **ovarian cancer** / **ovarian neoplasm** / **ultrasonography** / **menopause** / / **CA125** / **HE4** /

## Streszczenie

***Cel:*** *Celem pracy była zewnętrzna walidacja wybranych modeli prognostycznych: autorstwa grupy International Ovarian Tumor Analysis opartych na regresji logistycznej (LR1 i LR2, 2005) oraz innych popularnych modeli przeznaczonych do diagnostyki różnicowej guzów jajnika takich jak: model zaproponowany przez Timmerman'a i wsp. (1999), Alcazar'a i wsp., (2003), indeks ryzyka nowotworu (RMI – risk of malignancy index, 1990) oraz testu ROMA (risk of malignancy algorithm, 2009). Modele zostały porównane z subiektywną oceną ultrasonograficzną przeprowadzoną przez doświadczonego specjalistę oraz skalami diagnostycznymi utworzonymi w naszym ośrodku: indeksem sonomorfologicznym (SM, 2004) i indeksem waskularyzacji (SD, 2004). Użyteczność analizowanych modeli została oceniona w zależności od różnych cech kliniczno-patologicznych, między innymi w zależności od statusu menopauzalnego pacjentki.*

***Metodyka:*** *W badaniu poddano analizie 268 guzów przydatków, w tym 167 guzów niezłośliwych i 101 nowotworów złośliwych jajnika. Każdy z guzów został oceniony w odniesieniu do kryteriów diagnostycznych analizowanych testów. Przed operacją oznaczono również poziom markerów CA125 i HE4.*

***Wyniki:*** *W całej badanej populacji wszystkie modele predykcyjne wykazały podobną wartość diagnostyczną. Natomiast, stwierdzono istotne różnice pomiędzy testami w sytuacji gdy analizowano osobno pacjentki przed i po menopauzie. Największe pole pod krzywą ROC (AU-ROC - area under the ROC curve) w grupie pacjentek przed menopauzą uzyskały: subiektywna ocena ultrasonograficzna (0,931), model Alcazar'a (0,912) oraz LR1 (0,909). Natomiast w grupie kobiet po menopauzie największy AU-ROC uzyskały: model Timmerman'a (0,973), ROMA (0,951) i RMI (0,938).*

***Wnioski :*** *Status menopauzalny jest podstawowym czynnikiem determinującym użyteczność modelu predykcyjnego w diagnostyce różnicowej guzów przydatków. Wszystkie z badanych modeli uzyskały wartość diagnostyczną umożliwiającą stosunkowo dokładną diagnostykę przedoperacyjną guzów przydatków.*

Słowa kluczowe: **rak jajnika** / **guz jajnika** / **ultrasonografia** / **CA125** / **HE4** /

## Introduction

One of the most challenging current problems in gynecology is the appropriate differentiation of adnexal masses. Identification of malignant ovarian tumors versus benign neoplasms and functional lesions is crucial, because it determines the necessity of surgery, the pre-operative work-up, who should perform the surgery (a gynecological oncologist or a general gynecologist), and adequate timing in the operation room [1]. Menopausal status often determines the selection of appropriate diagnostic and treatment methods.

Gynecologists around the world have developed many prognostic models, ultrasonographic morphological scales, and other risk of malignancy calculators that are used for differential diagnosis of ovarian tumors. However, the plurality of diagnostic tests confirms their imperfections. Over 10 years ago, the International Ovarian Tumor Analysis (IOTA) Group started a project to improve our ability to differentiate between benign and malignant ovarian tumors. Several years of comprehensive and broad studies resulted in a number of predictive models. Among these models, the most important are 2 models based on logistic regression (LR1 and LR2) [2, 3]. These models were externally validated in studies supervised by IOTA members, which provided encouraging results [2]. However, certain features of ovarian cancer biology may differ between various populations. Moreover, the skills and experience of the examiner may influence the performance of diagnostic tools. Thus, additional independent external validations are needed for objective recommendation of IOTA logistic regression models for triaging of adnexal masses.

## The aim of the study

The aim of this study was to externally validate the diagnostic value of IOTA logistic regression models and other popular prognostic models of ovarian tumors. The diagnostic value of the analyzed models was compared to subjective ultrasonographic assessment and scales previously developed in our department as well as with selected popular scoring systems and predictive models [4, 5]. Another interesting question was whether menopausal status was a key factor that could affect differential diagnosis of adnexal masses. We tested whether menopausal status was a factor that affected the performance of the predictive models evaluated in our study.

## Materials and methods

The study group included 268 patients diagnosed with and treated for ovarian tumors between 2006 and 2012. The group was selected from all women referred to our clinic with adnexal masses. The main inclusion criterion was the ultrasonographic appearance of a tumor that could not be classified as either "certainly benign" or "certainly malignant" based on a subjective ultrasonographic assessment performed by an experienced


© Polskie Towarzystwo Ginekologiczne

ultrasonography specialist [6-8]. Tumors designated as "certainly benign" or "certainly malignant" were excluded, while the rest of the tumors were declared to be "suspicious" tumors and were included in our analysis. Subsequently, the "suspicious" tumors were again classified as either "benign" or "malignant" on the basis of the final subjective ultrasonography assessment. Thus, the tumors classified as "suspicious" during the first evaluation were subsequently classified as "benign" or "malignant." Tumors considered "probably benign" were subsequently classified as "benign," while tumors thought to be "probably malignant" were classified as "malignant." After ultrasonographic examination, the examiner was obligated to give his own subjective impression and to classify the tumor as either "benign" or "malignant." This discrimination was a subjective ultrasonographic assessment based on the knowledge and personal experience of the examiner. In general, the examiner judged unilocular and multilocular cysts without any papillary projections, even projections <3 mm, or without solid components, to be benign. In some cases, specific diagnoses were possible (e.g. endometrioma, teratoma) on the basis of pattern recognition on the gray-scale ultrasonography image; those tumors were classified as "certainly benign." Cystic tumors with solid components and more complex, irregular tumors were judged to be suspicious.

The study population includes the group of tumors that could not be easily determined as having malignant potential based on ultrasonographic examination. Characteristics of the analyzed patients are summarized in Table I.

Ultrasonographic examinations were performed 1 to 3 days before surgical treatment during the first 10 days of the menstrual cycle. The examination was performed by using an Aloka 3500 (Hitachi Aloka, Tokyo, Japan) with a 7.5 MHz endovaginal probe. A transabdominal probe was also used for large tumors. In patients with tumors on both sides, the larger and more complex tumor was considered for diagnosis. Tumors were ultrasonographically assessed according to the rules proposed in 2000 by the IOTA Group [9]. The structures of the analyzed tumors obtained by using ultrasonography are presented in Table II. Ultrasonographic examinations were performed by a single experienced examiner (R.M.) who recorded all features of the tumors required for development of the IOTA logistic regression models (LR1 and LR2), Timmerman's logistic regression model, the sonomorphological index (SM) with a cut-off at 8 points, the vascularization index (SD) with a cut-off at 4 points, Alcazar's index with a cut-off at 6 points, and the risk of malignancy index (RMI) with a cut-off at 200 points [3-5, 10-12]. The risk of ovarian malignancy algorithm (ROMA) was also calculated with cut-offs at 0.131 for premenopausal and 0.277 for postmenopausal patients [13, 14]. The originally developed RMI requires information about the presence of metastases [11, 15]. However, in our study, patients presenting with certain or highly suspected cancerous disease were excluded from the analysis; therefore, we provided a "0" for all patients after considering the RMI calculation.

Cancer antigen 125 (CA125) serum levels were assessed 1 to 5 days before the operation by using an immunoenzymatic test (ST AIA-PACK OVCA, Tosoh Bioscience, Tokyo, Japan). Human epididymis 4 (HE4) levels were assessed in stored serum samples (<−82 °C) that were obtained 1 to 5 days before surgery. An enzyme-linked immunosorbent assay (Fujirebio Diagnostics AB, Goteborg, Sweden) was used to evaluate HE4 levels.

All patients included in the study were referred for surgery via either a laparotomy or a laparoscopic approach. After subsequent histopathological examination, there were 167 benign ovarian tumors and 101 malignant ovarian tumors, including 14 tumors of borderline malignancy. The results of the histopathological examination are shown in Table 3. Some diagnoses are very common and can usually be easily recognized by using ultrasonography (e.g. endometrioid cysts, teratomas, and hemorrhagic cysts). However, according to our methodology, they meet the inclusion criteria because of their complex appearances on ultrasonography images.

Malignant tumors were classified according to the International Federation of Gynecology and Obstetrics (FIGO) disease stages as follows: Ia, 16 patients; Ib, 6 patients; Ic, 9 patients; IIa, 9 patients; IIb, 6 patients; IIc, 2 patients; IIIa, 10 patients; IIIb, 10 patients; IIIc, 33 patients. There were no patients diagnosed with stage IV disease because patients with metastatic lesions did not meet our inclusion criteria. The distribution of the histological grades of the analyzed malignant neoplasms was as follows: G1, 37 patients; G2, 31 patients; G3, 33 patients. The data was analyzed for the entire patient population, and also after stratification of pre- and postmenopausal patients.

Statistical evaluation was performed by using R software, version 2.15.3 (R Foundation for Statistical Computing, Vienna, Austria) [16]. We used the ROCR library (version 1.0–4) to develop receiver operating characteristic (ROC) curves [17]. The area under the ROC curve (AUC), standard error, and confidence intervals were calculated by using the pROC library, version 1.5.4 [18]. The DeLong method was used to assess the standard error and confidence intervals for the AUCs [19]. Additionally, the DeLong method was used to compare the AUCs of different tests.

A common problem with comparing the performance of diagnostic scales is the lack of a widely accepted tool that combines important parameters such as accuracy (acc), specificity (spec) and sensitivity (sens). Therefore, we propose a new method based on the notion of a t-norm, that is, a well-developed mathematical concept for aggregating numerical data (for more information, refer to [20, 21]). Our aim was to emphasize the role of sensitivity in medical diagnosis. We chose a popular product t-norm T, defined as:

$$T(a, b) = a \cdot b$$

We used this to construct a t-score method given by the following formula:

$$\begin{aligned} &\text{t-score (acc, spec, sens)} = T(T[acc, sens], T[spec, sens]) \\ &= T(acc \cdot sens, spec \cdot sens) \\ &= acc \cdot sens \cdot spec \cdot sens \\ &= acc \cdot sens2 \cdot spec \end{aligned}$$

Consider the following example, where we have 2 diagnostic scales: A (acc = 0.8, spec = 0.7, sens = 0.9) and B (acc = 0.9, spec = 0.9, sens = 0.7). It is difficult to identify the best scale without aggregation. By using the proposed t-score method, it is easy to see that scale A (t-score = 0.453) performs better than scale B (t-score = 0.397).

The local Ethics Committee approved this study.

## Results

Table 4 shows the results of AUC analysis of the predictive models and subjective ultrasonographic assessment in the whole study population. This table also shows reported AUCs from the


© Polskie Towarzystwo Ginekologiczne

**Table I.** Characteristics of analyzed patients.

| | | Benign ovarian tumors (n = 167) | | | Malignant ovarian tumors (n = 101) | | | |
|---|---|---|---|---|---|---|---|---|
| | | median | Range | | median | Range | | P - value |
| | | | minimal | maximal | | minimal | maximal | |
| **Age (years)** | | 40 | 15 | 74 | 53 | 21 | 84 | 0.00001 |
| **BMI** | | 22 | 17 | 42 | 25 | 18 | 50 | 0.00001 |
| **History of deliveries** | | 1 | 0 | 5 | 2 | 0 | 7 | 0.00011 |
| **Tumor volume (cm3)** | | 164.5 | 11 | 4187 | 484 | 14 | 4187 | 0.00006 |
| **CA125 (IU/ml)** | | 24 | 0.53 | 2367 | 303.9 | 5.71 | 4909 | 0.00001 |
| **HE4 (pmol/l)** | | 32.8 | 18.85 | 157 | 180 | 19.26 | 4246.6 | 0.00009 |
| | | Number (%) | | | Number (%) | | | |
| Menopausal status | Premenopausal | 131 (78%) | | | 46 (46%) | | | 0.00001 |
| | Postmenopausal | 36 (22%) | | | 55 (55%) | | | |

**Table II.** The structure of analyzed tumors as observed with ultrasonography.

| Ovarian tumor classification | Benign ovarian tumors | Malignant ovarian tumors |
|---|---|---|
| Unilocular cyst | 43 | 4 |
| Unilocular-solid tumor | 29 | 8 |
| Multilocular cyst | 35 | 7 |
| Multilocular-solid tumor | 48 | 56 |
| Solid tumor | 10 | 26 |
| Unclassified | 2 | 0 |

**Table III.** Results of histopathological examination.

| Benign ovarian tumor | Number |
|---|---|
| Endometrioid cyst | 55 |
| Corpus luteum cyst | 4 |
| Adult teratoma | 26 |
| Serous cystadenoma | 37 |
| Mucinous cystadenoma | 17 |
| Hemorrhagic cyst | 9 |
| Tubo-ovarian abscess | 6 |
| Adenofibroma | 1 |
| Theca cell tumor | 4 |
| Brenner tumor | 3 |
| Peduculated leiomyoma | 5 |
| **Malignant ovarian tumor** | |
| Serous adenocarcinoma | 38 |
| Mucinous adenocarcinoma | 5 |
| Endometrioid adenocarcinoma | 10 |
| Clear cell adenocarcinoma | 6 |
| Undifferentiated carcinoma | 19 |
| Granulosa cell tumor | 2 |
| Borderline ovarian tumor | 14 |
| Metastatic ovarian tumor | 7 |

original studies. The highest AUC was found for subjective ultrasonographic assessment. However, other predictive models also had high AUCs. Generally, we found only a few statistically significant differences in AUCs between the studied models. Subjective ultrasonographic assessment was superior to SM ($p = 0.027$) and the IOTA logistic regression model LR2 ($p = 0.014$). The predictive model developed by Timmerman et al. had a higher AUC than ROMA ($p = 0.044$) and RMI ($p = 0.023$). The differences among the other diagnostic tools were not statistically significant. The results of AUC comparison in the whole study population are summarized in Table V.

When the study group was subdivided into pre- and postmenopausal patients, there were many more differences in AUCs among the analyzed predictive models. In the subgroup of premenopausal patients, the highest AUC was achieved via subjective ultrasonographic assessment, and it was significantly higher than the AUCs for LR2 ($p = 0.01$), ROMA ($p = 0.003$), and RMI ($p = 0.002$). The IOTA logistic regression model LR1 and the scoring system developed by Alcazar also achieved high AUCs. The lowest AUCs were found for RMI and ROMA.

Surprisingly, the hierarchy of test utility in the subgroup of postmenopausal patients was the inverse of the hierarchy in premenopausal patients. The highest AUC in the postmenopausal patient group was reported for Timmerman's logistic regression model; it was higher than the AUC for subjective ultrasonographic assessment ($p = 0.02$). Comparisons of AUCs also showed that Timmerman's model was superior to Alcazar's scoring system ($p = 0.007$), LR1 ($p = 0.01$), LR2 ($p = 0.001$), SM ($p < 0.001$), and SD ($p = 0.02$). However, there were no differences in AUCs between Timmerman's model and RMI ($p = 0.058$) or ROMA ($p = 0.125$), which also achieved high AUCs.

The results of AUC evaluation and AUC comparisons in premenopausal and postmenopausal patients are shown in Table VI and Table VII, respectively.

Sensitivity, specificity, accuracy, negative and positive predictive values, and t-norm aggregation for the diagnostic models


© Polskie Towarzystwo Ginekologiczne

**Table IV.** Area under the curve (AUC) analysis of predictive models and subjective ultrasonographic assessment in the whole studied population.

| Method | Original report on test set | | Prospective testing on external validation data set | | Difference between original AUC and external AUC |
|---|---|---|---|---|---|
| | AUC | (95% CI) | AUC | (95% CI) | p-value |
| **LR1 [3]** | 0.936 | (0.916-0.956) | 0.914 | (0.879-0.949) | 0.4136 |
| **LR2 [3]** | 0.916 | (0.895-0.937) | 0.884 | (0.842-0.925) | 0.2813 |
| **Timmerman [12]** | 0.904 | (0.844-0.964) | 0.924 | (0.888-0.960) | 0.7495 |
| **SM [4]** | 0.883 | (0.870-0.896) | 0.887 | (0.846-0.928) | 0.8713 |
| **SD [5]** | 0.932 | (0.918-0.946) | 0.864 | (0.808-0.919) | 0.0298 |
| **Alcazar [10]** | 0.950 | (0.937-0.963) | 0.914 | (0.879-0.948) | 0.1049 |
| **RMI [11, 15]** | — | — | 0.898 | (0.855-0.942) | — |
| **ROMA [13, 14]** | — | — | 0.904 | (0.847-0.961) | — |
| **Sub [27]** | 0.963 | — | 0.927 | (0.895-0.959) | — |

LR1 – logistic regression model No. 1; LR2 – logistic regression model No. 2; Timmerman – Logistic regression model proposed by Timmerman in 1999, SM – sonomorphologic index; SD - vascularization index; Alcazar – scoring system proposed by J.L. Alcazar in 2003; RMI – risk of malignancy index; ROMA - risk of malignancy algorithm; Sub – subjective ultrasonographic assessment

**Table V.** The p-value of area under the ROC curve comparison in the whole group of studied patients.

| | Alcazar | LR1 | LR2 | Timmerman | SM | SD | ROMA | RMI | Sub |
|---|---|---|---|---|---|---|---|---|---|
| **Alcazar [10]** | x | | | | | | | | |
| **LR1 [3]** | 0.959 | x | | | | | | | |
| **LR2 [3]** | 0.062 | 0 | x | | | | | | |
| **Timmerman [12]** | 0.586 | 0.606 | 0.06 | X | | | | | |
| **SM [4]** | 0.172 | 0.089 | 0.844 | 0.134 | x | | | | |
| **SD [5]** | 0.077 | 0.981 | 0.262 | 0.095 | 0.257 | x | | | |
| **ROMA [13, 14]** | 0.802 | 0.39 | 0.778 | **0.044** | 0.899 | 0.6 | X | | |
| **RMI [11, 15]** | 0.362 | 0.326 | 0.612 | **0.023** | 0.622 | 0.525 | 0.124 | x | |
| **Sub [27]** | 0.441 | 0.382 | **0.014** | 0.873 | **0.027** | 0.287 | 0.134 | 0.143 | x |

LR1 – logistic regression model No. 1; LR2 – logistic regression model No. 2; Timmerman – Logistic regression model proposed by Timmerman in 1999, SM – sonomorphologic index; SD - vascularization index; Alcazar – scoring system proposed by J.L. Alcazar in 2003; RMI – risk of malignancy index; ROMA - risk of malignancy algorithm; Sub – subjective ultrasonographic assessment

and subjective ultrasonographic assessment are shown in Table VIII. This table also contains information about these parameters from the original reports. In our study, all of the diagnostic models were less accurate than was reported in the original studies. Additionally, Table 9 shows the test characteristics of the diagnostic models that were different in the premenopausal and postmenopausal patient groups (Table IX).

Interesting results are given by the proposed t-score measure. In the entire patient population, the best outcome for this measure was achieved via subjective ultrasonographic assessment (t-score = 0.602), followed by Timmerman's model (t-score = 0.509), ROMA (t-score = 0.506), and Alcazar's scoring system (t-score = 0.496). Among premenopausal patients, the best result was once again achieved with subjective ultrasonographic assessment (t-score = 0.574), followed by Alcazar's scoring system and LR1 (t-score = 0.51 and 0.46, respectively). Surprisingly, in postmenopausal patients, the best result was achieved with either Timmerman's model or ROMA, which performed equally well (t-score = 0.605), followed by subjective ultrasonographic assessment (t-score = 0.574).

## Discussion

External validation of predictive models used for differentiation of adnexal masses is essential to prove their practical utility. Our study confirmed that a validation by independent clinicians in a new study population could result in decreased diagnostic performance of the evaluated models [22-24]. However, both IOTA models (LR1 and LR2) achieved satisfactory accuracy.

This study found a strong relationship between model performance and the menopausal status of patients that was observed for all of the studied models. Our results are in opposition to the external validation performed by Van Holsbeke et al., where both IOTA models (LR1 and LR2) achieved very similar AUCs in premenopausal and postmenopausal patients [2]. Differences in the patient populations in these 2 studies may explain the different results. We narrowed our study population to only include patients who presented with tumors that were difficult to categorize according to subjective ultrasonographic assessment; there was no such limitation in the Van Holsbeke et al. study [2].

The most interesting and important research focuses on difficult-to-assess ovarian tumors, because the risk of false results is


© Polskie Towarzystwo Ginekologiczne

**Table VI.** Area under the curve (AUC) value for subjective ultrasonographic assessment and analyzed predictive models in subgroup of pre- and postmenopausal women

| Method | Premenopausal women | | Postmenopausal women | | Difference between premenopausal AUC and postmenopausal AUC |
|---|---|---|---|---|---|
| | AUC | (95% CI) | AUC | (95% CI) | p-value |
| LR1 [3] | 0.909 | (0.861-0.957) | 0.868 | (0.781-0.955) | 0.4178 |
| LR2 [3] | 0.876 | (0.817-0.935) | 0.831 | (0.735-0.926) | 0.4335 |
| Timmerman [12] | 0.882 | (0.817-0.946) | 0.973 | (0.949-0.998) | 0.0103 |
| SM [4] | 0.891 | (0.831-0.950) | 0.807 | (0.711-0.902) | 0.1437 |
| SD [5] | 0.868 | (0.795-0.941) | 0.823 | (0.702-0.944) | 0.5331 |
| Alcazar [10] | 0.912 | (0.861-0.964) | 0.894 | (0.833-0.955) | 0.6564 |
| RMI [11, 15] | 0.836 | (0.754-0.918) | 0.938 | (0.890-0.986) | 0.0350 |
| ROMA [13, 14] | 0.821 | (0.696-0.947) | 0.951 | (0.902-0.999) | 0.0585 |
| Sub [27] | 0.931 | (0.888-0.974) | 0.877 | (0.791-0.962) | 0.2636 |

LR1 – logistic regression model No. 1; LR2 – logistic regression model No. 2; Timmerman – Logistic regression model proposed by Timmerman in 1999, SM – sonomorphologic index; SD – vascularization index; Alcazar – scoring system proposed by J.L. Alcazar in 2003; RMI – risk of malignancy index; ROMA – risk of malignancy algorithm; Sub – subjective ultrasonographic assessment

**Table VII.** The p-value of area under the curve comparison in the group of premenopausal and postmenopausal women

| Premenopause | Alcazar | LR1 | LR2 | Timmerman | SM | SD | ROMA | RMI | sub |
|---|---|---|---|---|---|---|---|---|---|
| Alcazar [10] | x | | | | | | | | |
| LR1 [3] | 0.83 | x | | | | | | | |
| LR2 [3] | 0.055 | **0.01** | x | | | | | | |
| Timmerman [12] | 0.302 | 0.35 | 0.869 | x | | | | | |
| SM [4] | 0.473 | 0.479 | 0.584 | 0.82 | x | | | | |
| SD [5] | 0.197 | 0.742 | 0.326 | 0.897 | 0.338 | x | | | |
| ROMA [13, 14] | 0.1 | 0.078 | 0.146 | 0.11 | 0.125 | 0.573 | x | | |
| RMI [11, 15] | **0.005** | **0.015** | 0.155 | **0.043** | 0.102 | 0.573 | 0.188 | x | |
| Sub [27] | 0.409 | 0.221 | **0.01** | 0.108 | 0.121 | 0.393 | **0.003** | **0.002** | x |
| **Postmenopause** | | | | | | | | | |
| Alcazar [10] | x | | | | | | | | |
| LR1 [3] | 0.443 | x | | | | | | | |
| LR2 [3] | 0.092 | 0.068 | x | | | | | | |
| Timmerman [12] | **0.007** | **0.01** | **0.001** | x | | | | | |
| SM [4] | **0.043** | 0.109 | 0.556 | **<0.001** | x | | | | |
| SD [5] | 0.116 | 0.818 | 0.364 | **0.02** | 0.254 | x | | | |
| ROMA [13, 14] | 0.358 | 0.368 | 0.064 | 0.125 | 0.111 | 0.44 | x | | |
| RMI [11, 15] | 0.23 | 0.107 | **0.024** | 0.058 | **0.011** | 0.208 | 0.819 | x | |
| Sub [27] | 0.691 | 0.839 | 0.353 | **0.02** | 0.099 | 0.741 | 0.848 | 0.3 | x |

LR1 – logistic regression model No. 1; LR2 – logistic regression model No. 2; Timmerman – Logistic regression model proposed by Timmerman in 1999, SM – sonomorphologic index; SD – vascularization index; Alcazar – scoring system proposed by J.L. Alcazar in 2003; RMI – risk of malignancy index; ROMA – risk of malignancy algorithm; Sub – subjective ultrasonographic assessment

high in this group. The diagnosis of advanced-stage disease and the subsequent decision to operate at an oncology center is clear. Similarly, a large group of tumors (e.g. simple cysts, endometrioid tumors, and dermoid cysts) is easy to diagnose as benign. The risk of malignancy in this group is extremely low. Therefore, the most important and interesting tumors are those that pose problems in ultrasonography evaluation. According to the IOTA Group's recent publication by Valentin et al., only 7% to 10% of masses were suspicious and difficult to classify [6]. Our study analyzed this group of tumors as well as tumors that were "probably malignant" or "probably benign," where there was a degree of uncertainty. We think this is likely a reason for the differences between the Van Holsbeke et al. study and our own. Furthermore, these inclusion criteria may be a reason for the decreased prognostic values of the analyzed diagnostic models compared to the original studies.


© Polskie Towarzystwo Ginekologiczne

**Table VIII.** Accuracy, sensitivity, specificity, positive and negative predictive values and t-score for diagnostic models and subjective ultrasonographic assessment in the original report and prospective analysis.

| Method | Original report on test set | | | | | Prospective testing | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | SENS | SPEC | PPV | NPV | ACC | SENS | SPEC | PPV | NPV | t-score |
| LR1 [3] | 0.798 | 0.933 | 0.755 | 0.554 | 97.3 | 0.761 | 0.960 | 0.641 | 0.618 | 0.964 | 0.450 |
| LR2 [3] | | 0.890 | 0.730 | | | 0.765 | 0.950 | 0.653 | 0.623 | 0.956 | 0.451 |
| Timmerman [12] | | 0.857 | 0.811 | 0.632 | 0.938 | 0.869 | 0.802 | 0.910 | 0.844 | 0.884 | 0.509 |
| SM [4] | 0.806 | 0.867 | 0.770 | 0.691 | 0.907 | 0.784 | 0.921 | 0.701 | 0.650 | 0.936 | 0.465 |
| SD [5] | 0.91 | 0.867 | 0.933 | 0.877 | 0.927 | 0.741 | 0.582 | 0.916 | 0.883 | 0.667 | 0.23 |
| Alcazar [10] | 0.967 | 1.0 | 0.949 | 0.912 | 1.0 | 0.843 | 0.832 | 0.850 | 0.771 | 0.893 | 0.496 |
| RMI [11, 15] | | 0.85 | 0.97 | | | 0.834 | 0.781 | 0.865 | 0.773 | 0.870 | 0.440 |
| ROMA [13, 14] | | 0.887 | 0.747 | 0.601 | 0.939 | 0.865 | 0.804 | 0.906 | 0.849 | 0.875 | 0.506 |
| Sub [27] | | 0.902 | 0.929 | | | 0.892 | 0.861 | 0.910 | 0.853 | 0.916 | 0.602 |

LR1 – logistic regression model No. 1; LR2 – logistic regression model No. 2; Timmerman – Logistic regression model proposed by Timmerman in 1999, SM – sonomorphologic index; SD – vascularization index; Alcazar – scoring system proposed by J.L. Alcazar in 2003; RMI – risk of malignancy index; ROMA – risk of malignancy algorithm; Sub – subjective ultrasonographic assessment

**Table IX.** Accuracy, sensitivity, specificity, positive and negative predictive values and t-scores in premenopausal and postmenopausal women.

| Method | Premenopausal women | | | | | | Postmenopausal women | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | SENS | SPC | PPV | NPV | t-score | ACC | SENS | SPC | PPV | NPV | t-score |
| LR1 [3] | 0.768 | 0.913 | 0.718 | 0.532 | 0.959 | 0.46 | 0.747 | 1 | 0.361 | 0.705 | 1 | 0.27 |
| LR2 [3] | 0.78 | 0.891 | 0.74 | 0.547 | 0.951 | 0.459 | 0.736 | 1 | 0.333 | 0.696 | 1 | 0.245 |
| Timmerman [12] | 0.864 | 0.674 | 0.931 | 0.775 | 0.891 | 0.366 | 0.879 | 0.909 | 0.833 | 0.893 | 0.857 | 0.605 |
| SM [4] | 0.785 | 0.87 | 0.756 | 0.556 | 0.943 | 0.449 | 0.78 | 0.964 | 0.5 | 0.746 | 0.9 | 0.362 |
| SD [5] | 0.794 | 0.585 | 0.934 | 0.857 | 0.77 | 0.254 | 0.667 | 0.58 | 0.864 | 0.906 | 0.475 | 0.194 |
| Alcazar [10] | 0.859 | 0.826 | 0.87 | 0.691 | 0.934 | 0.51 | 0.813 | 0.836 | 0.778 | 0.852 | 0.757 | 0.442 |
| RMI [11, 15] | 0.842 | 0.628 | 0.914 | 0.711 | 0.88 | 0.303 | 0.818 | 0.906 | 0.686 | 0.814 | 0.828 | 0.46 |
| ROMA [13, 14] | 0.855 | 0.6 | 0.937 | 0.75 | 0.881 | 0.288 | 0.879 | 0.917 | 0.818 | 0.892 | 0.857 | 0.605 |
| Sub [27] | 0.904 | 0.826 | 0.931 | 0.809 | 0.938 | 0.574 | 0.868 | 0.891 | 0.833 | 0.891 | 0.833 | 0.574 |

LR1 – logistic regression model No. 1; LR2 – logistic regression model No. 2; Timmerman – Logistic regression model proposed by Timmerman in 1999, SM – sonomorphologic index; SD – vascularization index; Alcazar – scoring system proposed by J.L. Alcazar in 2003; RMI – risk of malignancy index; ROMA – risk of malignancy algorithm; Sub – subjective ultrasonographic assessment

We found the highest AUCs in the postmenopausal group of patients for tests incorporating biomarker assessment (RMI, ROMA, and Timmerman's logistic regression model). Recent studies have shown that neither CA125 nor HE4 improved the diagnostic performance of subjective ultrasonographic assessment [6, 8, 25, 26]. However, our results suggest that evaluation of biomarker levels within an ultrasonographic model results in significantly higher diagnostic utility compared to subjective ultrasonographic assessment in the postmenopausal patient group. We have a few hypotheses that may explain the differences between studies. First of all, our study used a combination of biomarker assessment (CA125 and HE4 in the case of ROMA) or a combination of biomarkers with an ultrasonographic scoring system or predictive model (RMI, Timmerman's model), while previous papers assessed a single marker [8, 25]. Secondly, in our previous work we did not stratify patients according to the menopausal status [8]. Thirdly, although we work in a tertiary gynecological center that specializes in ovarian cancer treatment, our ultrasonographic experience is probably less than the authors mentioned in the Valentin et al., study [25].

The hierarchy of test performance in premenopausal patients was almost completely the inverse of the hierarchy in postmenopausal patients. In the group of premenopausal patients, the highest AUC was achieved with subjective ultrasonographic assessment. This group of patients was characterized by the frequent presence of functional ovarian cysts, endometrioid cysts, and adult-type teratomas. The ultrasonographic features of these tumors were characteristic. Thus, these tumors can usually be correctly classified by an experienced ultrasonography specialist by using pattern recognition. However, sometimes their morphology can be complex, and this may affect the specificity of predictive models [27]. This was confirmed by our study, which only ana-


© Polskie Towarzystwo Ginekologiczne

lyzed complex tumors. Furthermore, endometrioid cysts may increase the levels of CA125 and HE4. This may be responsible for the lower diagnostic utility of predictive models that incorporate biomarker assessment in premenopausal women.

We have proposed t-score evaluation as a new tool for the assessment of diagnostic test utility. Our evaluation of t-scores confirmed the previous finding that, for premenopausal patients, it is better to rely on an experienced ultrasonography specialist rather than mathematical models. Alternatively, among postmenopausal patients, Timmerman's model and ROMA achieved better overall results, though the difference compared to subjective ultrasonographic assessment was not overwhelming. In conclusion, our study shows that:

1. Menopausal status is a key factor that impacts the utility of prognostic models for differential diagnosis of ovarian tumors;
2. IOTA logistic regression models (LR1 and LR2) are reasonable diagnostic tools for less experienced ultrasonography specialists, especially in the premenopausal patient group;
3. Ultrasonographic scoring systems and predictive models that incorporate biomarker assessment are potent diagnostic tools for differentiation of adnexal masses in postmenopausal patients.

**Oświadczenie autorów:**

1. Rafał Moszyński – autor koncepcji i założeń pracy, wykonanie badań ultrasonograficznych – autor odpowiedzialny za manuskrypt.
2. Patryk Zywica – autor obliczeń matematycznych, analiza statystyczna wyników, przygotowanie manuskryptu.
3. Andrzej Wojtowicz – autor obliczeń matematycznych, analiza statystyczna wyników, przygotowanie manuskryptu.
4. Sebastian Szubert – przygotowanie manuskryptu, autor zgłaszający, interpretacja wyników, zbieranie materiału.
5. Stefan Sajdak – analiza i interpretacja wyników, korekta i akceptacja ostatecznego kształtu manuskryptu.
6. Anna Stachowiak – współautorka tekstu pracy, analiza uzyskanych wyników.
7. Krzysztof Dyczkowski – nadzór i korekta obliczeń matematycznych i statystycznych, opracowanie wyników, korekta i akceptacja ostatecznego kształtu manuskryptu.
8. Maciej Wygralak – analiza i interpretacja wyników, korekta i akceptacja ostatecznego kształtu manuskryptu.
9. Dariusz Szpurek – analiza i interpretacja wyników, korekta i akceptacja ostatecznego kształtu manuskryptu.

**Źródło finansowania:**

Praca nie była finansowana przez żadną instytucję naukowo-badawczą, stowarzyszenie ani inny podmiot, autorzy nie otrzymali żadnego grantu.

**Konflikt interesów:**

Autorzy nie zgłaszają konfliktu interesów oraz nie otrzymali żadnego wynagrodzenia związanego z powstawaniem pracy.

## 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, 422-436.
2. Van Holsbeke C, Van Calster B, Bourne T, [et al.]. External validation of diagnostic models to estimate the risk of malignancy in adnexal masses. Clin Cancer Res. 2012, 18, 815-825.
3. Timmerman D, Testa AC, Bourne T, [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, 8794-8801.
4. Szpurek D, Moszynski R, Zietkowiak W, [et al.]. An ultrasonographic morphological index for prediction of ovarian tumor malignancy. *Eur J Gynaecol Oncol.* 2005, 26, 51-54.
5. Szpurek D, Moszynski R, Sajdak S. Clinical value of the ultrasound Doppler index in determination of ovarian tumor malignancy. *Eur J Gynaecol Oncol.* 2004, 25, 442-444.
6. 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, 456-465.
7. Timmerman D, Schwarzler P, Collins WP, [et al.]. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet Gynecol.* 1999, 13, 11-16.
8. Moszynski R, Szubert S, Szpurek D, [et al.]. Usefulness of the HE4 biomarker as a second-line test in the assessment of suspicious ovarian tumors. *Arch Gynecol Obstet.* 2013, 288 96), 1377-1383. doi: 10.1007/s00404-013-2901-1
9. Timmerman D, Valentin L, Bourne TH, [et al.]. 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.
10. Alcazar JL, Merce LT, Laparte C, [et al.]. A new scoring system to differentiate benign from malignant adnexal masses. Am J Obstet Gynecol. 2003, 188, 685-692.
11. Jacobs I, Oram D, Fairbanks J, [et al.]. 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, 922-929.
12. Timmerman D, Bourne TH, Tailor A, [et al.]. A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model. *Am J Obstet Gynecol.* 1999, 181,57-65.
13. Montagnana M, Danese E, Ruzzenente O, [et al.]. The ROMA (Risk of Ovarian Malignancy Algorithm) for estimating the risk of epithelial ovarian cancer in women presenting with pelvic mass: is it really useful? *Clin Chem Lab Med.* 2011, 49, 521-525.
14. Moore RG, McMeekin DS, Brown AK, [et al.]. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol Oncol.* 2009, 112, 40-46.
15. Tingulstad S, Hagen B, Skjeldestad FE, [et al.]. Evaluation of a risk of malignancy index based on serum CA125, ultrasound findings and menopausal status in the pre-operative diagnosis of pelvic masses. *Br J Obstet Gynaecol.* 1996, 103, 826-831.
16. The R Development Core Team. A language and environment for statistical computing. Vienna, Austria: *R Foundation for Statistical Computing,* 2012.
17. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics.* 2005, 21, 3940-3941.
18. 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, 77.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics.* 1988, 44, 837-845.
20. Klement EP, Mesiar R, Pap E. Triangular norms. Dordrecht ; Boston: *Kluwer Academic Publishers,* 2000.
21. Wygralak M. Intelligent counting under information imprecision : applications to intelligent systems and decision support. New York: *Springer;* 2012.
22. Mol BW, Boll D, De Kanter M, [et al.]. Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. *Gynecol Oncol.* 2001, 80, 162-167.
23. Timmerman D, Verrelst H, Collins WP, Re: Mol [et al.]. Distinguishing the benign and malignant adnexal mass: an external validation of prognostic models. *Gynecol Oncol.* 2001, 80, 162-167. *Gynecol Oncol.* 2001, 83, 66-168.
24. Mol BW, Boll D, Sijmons EA, Brolmann HA. Reply: To the Editor:. *Gynecol Oncol.* 2001, 83, 167-168.
25. 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, 345-354.
26. Terzic M, Dotlic J, Brndusic N, Arsenovic N, Likic I, Ladjevic N, et al. Histopathological diagnoses of adnexal masses: which parameters are relevant in preoperative assessment? *Ginekol Pol.* 201, 84, 7030-708.
27. Valentin L. Use of morphology to characterize and manage common adnexal masses. *Best Pract Res Clin Obstet Gynaecol.* 2004, 18, 71-89.
28. 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, 684-691.


© Polskie Towarzystwo Ginekologiczne