Comparison of O-RADS, GI-RADS, and IOTA simple rules regarding malignancy rate, validity, and reliability for diagnosis of adnexal masses

Mohammad Abd Alkhalik Basha, Maha Ibrahime Metwally, Shrif A. Gamil, Hamada M. Khater, Sameh Abdelaziz Aly, Ahmed A. El Sammak, et al.

European Radiology

ISSN 0938-7994
Volume 31
Number 2

Eur Radiol (2021) 31:674-684
DOI 10.1007/s00330-020-07143-7

Springer

European Radiology (2021) 31:674–684
https://doi.org/10.1007/s00330-020-07143-7

ULTRASOUND

Comparison of O-RADS, GI-RADS, and IOTA simple rules regarding malignancy rate, validity, and reliability for diagnosis of adnexal masses

Mohammad Abd Alkhalik Basha[1] · Maha Ibrahime Metwally[1] · Shrif A. Gamil[2] · Hamada M. Khater[3] · Sameh Abdelaziz Aly[3] · Ahmed A. El Sammak[1] · Mohamed M. A. Zaitoun[1] · Enass M. Khattab[1] · Taghreed M. Azmy[1] · Nader Ali Alayouty[1] · Nesreen Mohey[1] · Hosam Nabil Almassry[1] · Hala Y. Yousef[1] · Safaa A. Ibrahim[4] · Ekramy A. Mohamed[4] · Abd El Motaleb Mohamed[5] · Amira Hamed Mohamed Afifi[6] · Ola A. Harb[7] · Hesham Youssef Algazzar[3]

Received: 2 April 2020 / Revised: 27 May 2020 / Accepted: 3 August 2020 / Published online: 18 August 2020

Abstract

Objective The American College of Radiology (ACR) recently published the ovarian-adnexal reporting and data system (O-RADS) to provide guidelines to physicians who interpret ultrasound (US) examinations of adnexal masses (AM). This study aimed to compare the O-RADS with two other well-established US classification systems for diagnosis of AM.

Methods This retrospective multicenter study between May 2016 and December 2019 assessed consecutive women with AM detected by the US. Five experienced consultant radiologists independently categorized each AM according to O-RADS, gynecologic imaging reporting and data system (GI-RADS), and international ovarian tumor analysis (IOTA) simple rules. Pathology and adequate follow-up were used as reference standards for calculating the validity of three US classification systems for diagnosis of AM. Kappa statistics were used to assess the inter-reviewer agreement (IRA).

Results A total of 609 women (mean age, 48 ± 13.7 years; range, 18–72 years) with 647 AM were included. Of the 647 AM, 178 were malignant and 469 were benign. Malignancy rates were comparable to recommended rates by previous literature in O-RADS and IOTA, but higher in GI-RADS. O-RADS had significantly higher sensitivity for malignancy than GI-RAD and IOTA ($p = 0.003$ and 0.0007, respectively), but non-significant slightly lower specificity ($p > 0.05$). O-RADS, GI-RADS, and IOTA showed similar overall IRA ($\kappa = 0.77$, 0.69, and 0.63, respectively) with a tendency toward higher IRA with O-RADS than with GI-RADS and IOTA.

Conclusions O-RADS compares favorably with GI-RADS and IOTA. O-RADS had higher sensitivity than GI-RADS and IOTA simple rules with relatively similar specificity and reliability.

Key Points

- *The malignancy rates were comparable to recommended rates by previous literature in O-RADS and IOTA, but higher in GI-RADS.*
- *The O-RADS had significantly higher sensitivity for malignancy than GI-RADS and IOTA (96.8% vs 92.7% and 92.1%; $p = 0.003$ and 0.0007, respectively), but non-significant slightly lower specificity (92.8% vs 93.6% and 93.2%, respectively; $p > 0.05$).*
- *The O-RADS, GI-RADS, and IOTA showed similar overall inter-reviewer agreement (IRA) ($\kappa = 0.77$, 0.69, and 0.63, respectively), with a tendency toward higher IRA with O-RADS than with GI-RADS and IOTA.*

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-020-07143-7) contains supplementary material, which is available to authorized users.

✉ Mohammad Abd Alkhalik Basha
Mohammad_basha76@yahoo.com

[1] Department of Radio-diagnosis, Zagazig University, Zagazig, Egypt

[2] Department of Radio-diagnosis, Al-Ahrar Teaching Hospital, Zagazig, Egypt

[3] Department of Radio-diagnosis, Benha University, Benha, Egypt

[4] Department of Obstetrics & Gynecology, Zagazig University, Zagazig, Egypt

[5] Department of Clinical Oncology, Zagazig University, Zagazig, Egypt

[6] Department of Clinical Pathology, Zagazig University, Zagazig, Egypt

[7] Department of Pathology, Zagazig University, Zagazig, Egypt

Keywords Adnexal diseases · Ultrasonography · Data systems · Sensitivity and specificity · Reproducibility of results

Abbreviations

AM	Adnexal masses
AUC	Area under the curve
FIGO	Federation of Gynaecology and Obstetrics
GI-RADS	Gynecologic Imaging Reporting and Data System
IOTA	International Ovarian Tumor Analysis
IRA	Inter-reviewer agreement
O-RADS	Ovarian-Adnexal Reporting and Data System
ROC	Receiver operating characteristic
US	Ultrasound

Introduction

Ultrasound (US) continues to be the initial imaging modality of choice for the identification and characterization of adnexal masses (AM) [1, 2]. Structured reporting of AM findings was identified by a Society of Radiologists in Ultrasound consensus working group as a target for the investigation to improve the management of women with AM [3]. To date, many established guidelines and structured reporting have been developed using sonography to characterize AM, including subjective assessment, simple scoring systems, and statistically derived scoring systems [4–13].

In 2008, the International Ovarian Tumor Analysis (IOTA) group [5] proposed the use of US simple rules for the diagnosis of ovarian malignancy. These are based on a set of five US features indicative of a benign tumor (B features), and five US features indicative of a malignant tumor (M features). In 2009, Amor et al [9] designed the Gynecology Imaging Reporting and Data System (GI-RADS) as an attempt to allow standardized reporting of AM. This system is based on recognition patterns and criteria provided by the IOTA. Recently, the American College of Radiology (ACR) [12] published the Ovarian-Adnexal Reporting and Data System (O-RADS), which provides an up-to-date suggestion to stratify the AM according to sonographic features. The O-RADS offers a comprehensive algorithm that categorizes AM by their possibility of being normal (O-RADS 1), to high risk of malignancy (O-RADS 5) [13].

For the application of the US classification system for AM in clinical settings, it is essential to evaluate their validity and reproducibility. Several studies have investigated the validity of these risk stratification systems in the assessment of AM. However, data on the comparability and reproducibility of the systems, particularly from different readers, are limited. The purposes of this study, therefore, were to compare the O-RADS with GI-RADS, and IOTA simple rules regarding malignancy rate, validity, and inter-reviewer agreement (IRA) for diagnosis of AM.

Methods

Ethics approval

This multicenter retrospective cohort study was approved by the institutional review boards of the three participating institutions, and informed patient consent was obtained.

Study population

Between May 2016 and December 2019, the hospital databases of the three participating institutions were searched for women who were referred for clinically suspected AM. Initially, 902 consecutively registered women were identified. The hospital's electronic medical case records and case notes were reviewed for demographic data of the patients, such as age, menopausal status, clinical examinations, tumor marker levels, surgical findings, pathologic diagnosis, and follow-up. The selection criteria were as follows: (a) women with an adequate database; (b) women who underwent transabdominal or transvaginal US, or both; and (c) women with pathologic diagnosis or adequate follow-up. Exclusion criteria are listed in Fig. 1. Finally, 609 women were enrolled in this study. Two hundred fifty-eight of the 609 women have been previously reported [14]. This prior article evaluated the diagnostic performance and IRA of the GI-RADS for diagnosis of AM by US, whereas, in this manuscript, we compared the O-RADS with GI-RADS and IOTA regarding their validity and reliability for the diagnosis of AM.

Ultrasound examination

The following US machines were used for US examinations: Logiq 9, GE Healthcare, ClearVue 650, Philips Healthcare, GE Voluson S8 BT18, and SonoScape S40. Highly experienced radiologists (with over 10 years of pelvic US experience and had performed > 1000 US examinations per year) performed all US examinations. All women underwent either transvaginal (TV) or transabdominal US examination, or both. The transabdominal US was performed for virgin patients (n = 66) or those with large tumors that cannot be completely seen by the TV route (n = 71). The most important b-mode parameters (such as gain, frequency, number of foci and their depth, etc.) of the US machines were manually adjusted to obtain similar image impressions. The radiologists reported the following morphological features for each examined AM:

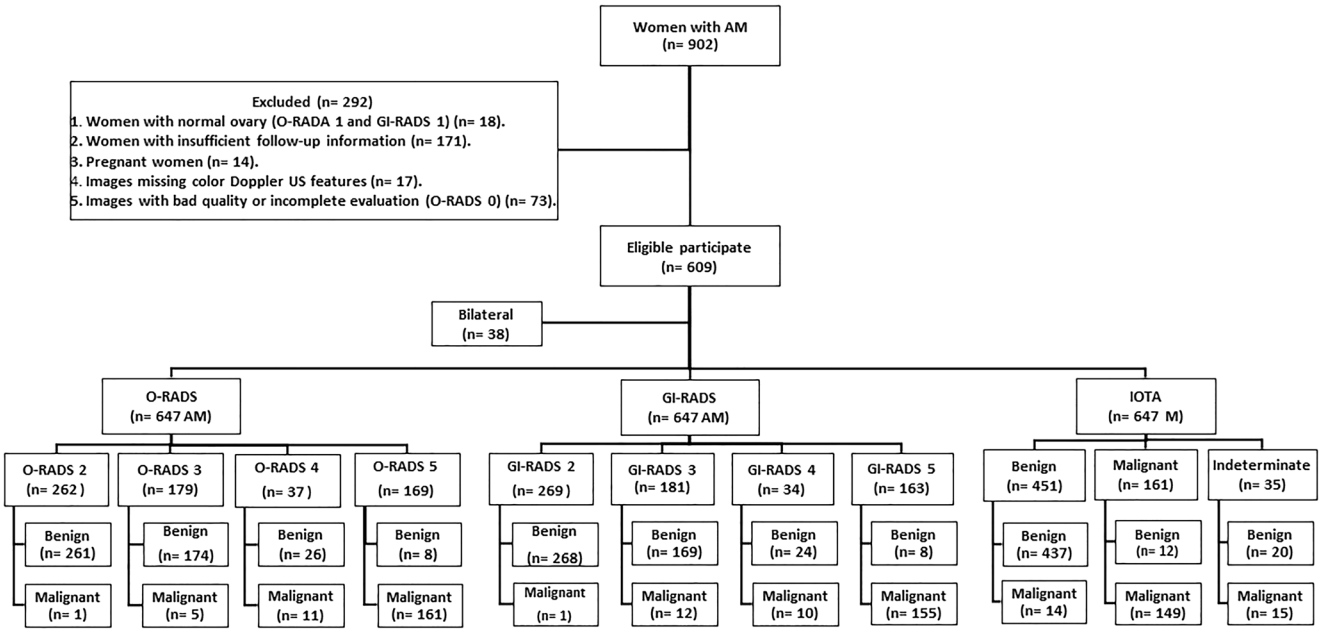

Fig. 1 Flow diagram of our study. AM = adnexal masses; O-RADS = Ovarian-Adnexal Reporting and Data System; GI-RADS = Gynecologic Imaging Reporting and Data System; IOTA = International Ovarian Tumor Analysis

laterality (unilateral or bilateral), maximum diameter, echogenicity, wall thickness, cystic content, solid areas, septations, solid papillary projections, the presence of ascites or peritoneal implant, and the pattern and the score of color Doppler.

Image analysis

The static US images of all examined women were collected in a central reading site where they were independently reviewed by five consultant radiologists from the three participating institutions but did not participate in the image capture. The radiologists had over 15 years of experience in pelvic imaging. At the time of their reviews, the radiologists had access to the original US reports but were blinded to the patients' medical information and pathologic results. In a first step, prior to starting the image analysis, the radiologists received 6 h of practical and theoretical training that described in detail the basic consensus on the lexicon of the three US classification systems. Afterward, the radiologists reviewed and analyzed the morphological features of each AM, and independently categorized the US images of each AM according to the O-RADS published by ACR [12, 13], GI-RADS designed by Amor et al [9, 10], and IOTA simple rules based on the descriptions proposed by Timmerman et al [5]. Lastly, a collaborative consensus reviewing with the contribution of the five radiologists was achieved to reach the final categorization of AM by each US classification system. In case of disagreement between radiologists, all parameters were discussed in detail until a final agreement was reached. The results of consensus reviewing were used to calculate the validity of each US classification system.

Reference standard

The definite diagnoses of AM were preferentially established based on the following:

(a) Postoperative pathological findings. Two experienced pathologists checked all specimens, and the results were obtained by consensus. The pathologist was blind to the patient's medical data and US findings. Histopathology of AM was classified according to the criteria recommended by the International Federation of Gynaecology and Obstetrics (FIGO) [15]. For statistical analysis, borderline AM were classified as malignant tumors.
(b) Adequate follow-up by regular US examinations every 3 months for more than 2 years after the initial US.

Statistical analysis

MedcCalc version 15.8 or SPSS version 26 was used for the analysis of collected data. Continuous variables were presented as means and standard deviations. Categorical variables were presented as numbers and percentages. Categorical variables were compared using the chi-square test or the Fisher exact test, while continuous variables were compared using the independent-samples *t* test. The fourfold table test was used to evaluate the validity of three US classification systems for the diagnosis of malignant AM using histopathology and

adequate follow-up as standard references. The receiver operating characteristic (ROC) curve was applied to determine the best cutoff values, calculate the areas under the curve (AUC), and comparatively analyze the validity of the three US classification systems. Weighted kappa (κ) statistics with 95% confidence intervals (CI) were used to evaluate overall IRA and between single reviewers for three US classification systems. The resulting κ values were interpreted as follows: poor agreement = 0.00–0.20, fair agreement = 0.21–0.40, moderate agreement = 0.41–0.60, good agreement = 0.61–0.80, and very good agreement = 0.81–1.00. The level of statistical significance was set at a p value of < 0.05.

Results

Patients and adnexal masses

During the study period, data on 609 women with at least one AM on US examination were collected. A total of 647 AM from 609 women (38 women (6.2%) had bilateral AM) were included in our final analysis. Figure 1 illustrates the flow diagram of our study. The clinical-pathologic characteristics of the included women (mean age, 48 ± 13.7 years; range, 18–72 years) are described in Table 1 provided in the Electronic Supplementary Materials. Four hundred ninety-nine AM were pathologically diagnosed. The remaining 148 AM (65 hemorrhagic cysts, 49 follicular cysts, 30 corpus luteum cysts, and 4 tubo-ovarian abscesses) were resolved spontaneously or after conservative medical treatment during follow-up and were considered to be benign. One hundred seventy-six (28.9%) women were postmenopausal, and 433 (71.1%) were premenopausal. Malignant AM were more common in postmenopausal women (71.8%). Of 647 AM, 178 (27.5%) were malignant and 469 (72.5%) were benign. Table 1 listed the final diagnoses of 647 AM. The most frequent benign AM was hemorrhagic cyst (20.5%), while the most frequent malignant AM was serous cystadenocarcinoma (29.8%). Three hundred eighty-seven AM needed conference consensus to reach the final categorization of AM. The disagreement between radiologists was highly reported in the GI-RADS classification system.

Table 1 Final diagnosis of 647 adnexal masses

Pathologic diagnosis	No. (%)
Benign adnexal masses	469 (72.5)
Follicular cyst >3 cm	55 (11.7)
Corpus luteum cyst >3 cm	49 (10.4)
Theca lutein cyst	6 (1.3)
Hemorrhagic cyst	96 (20.5)
Endometrioma	71 (15.1)
Dermoid cysts	68 (14.5)
Paraovarian cyst	24 (5.1)
Peritoneal inclusion cyst	23 (4.9)
Hydrosalpinx/tubo-ovarian abscess	28 (6)
Serous cystadenoma	25 (5.3)
Mucinous cystadenoma	19 (4.1)
Fibroma	5 (1.1)
Malignant adnexal masses	178 (27.5)
Serous cystadenocarcinoma	53 (29.8)
Mucinous cystadenocarcinoma	34 (19.1)
Borderline serous cystadenoma	20 (11.2)
Borderline mucinous cystadenoma	12 (6.7)
Germ cell tumor	21 (11.8)
Metastatic carcinoma	12 (6.7)
Malignant stromal tumors	11 (6.2)
Immature teratoma	7 (3.9)
Endometrioid carcinoma	6 (3.4)
Granulosa cell tumor	2 (1.1)

Distribution of categories in three US classification systems

The frequency of O-RADS, GI-RADS, and IOTA categories stratified by the system and reviewer is presented in Table 2.

Malignancy rates in the categories of three US classification systems

The malignancy rates of three US classification systems are presented in Table 3. Based on the O-RADS categories, the percentages of malignancy in O-RADS 2, 3, 4, and 5 were 0.4%, 2.8%, 30.6%, and 95.3%, respectively; the differences were statistically significant ($p < 0.001$). Based on the GI-RADS categories, the percentages of malignancy in GI-RADS 2, 3, 4, and 5 were 0.4%, 6.6%, 31.3%, and 95.1%, respectively; the differences were statistically significant ($p < 0.001$). Based on the IOTA simple rules, the percentages of malignancy in benign, malignant, and indeterminate AM were 3.1%, 92.5%, and 42.9%, respectively; the differences were statistically significant ($p < 0.001$).

Diagnostic validity of three US classification systems

The ROC curve analysis demonstrated that, for the O-RADS, the best cutoff value was > O-RADS 3. For the GI-RADS, the best cutoff value was > GI-RADS 3. For IOTA, the best cutoff value was the presence of only M features, no features, or the presence of both M and B features. Table 4 summarizes the per-lesion validity of the three US classification systems for the diagnosis of malignant AM using the consensus data. Considering combined O-RADS 4 and O-RADS 5 as a predictor for malignancy,

Table 2 Frequency distributions of O-RADS, GI-RADS, and IOTA for 647 adnexal masses stratified by system and reviewer

Category	Characteristic	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Consensus reviewing
O-RADS							
O-RADS 2	Almost certainly benign	276 (42.7)	285 (44)	249 (38.5)	217 (33.5)	232 (35.9)	262 (40.5)
O-RADS 3	Low risk	205 (31.7)	139 (21.5)	166 (25.7)	198 (30.6)	183 (28.3)	179 (27.7)
O-RADS 4	Intermediate risk	47 (7.3)	73 (11.3)	61 (9.4)	26 (4.1)	41 (6.3)	37 (5.7)
O-RADS 5	High risk	119 (18.3)	150 (23.2)	171 (26.4)	206 (31.8)	191 (29.5)	169 (26.1)
GI-RADS							
GI-RADS 2	Very probably benign	295 (45.6)	242 (37.4)	209 (32.3)	233 (36)	301 (46.5)	269 (41.6)
GI-RADS 3	Probably benign	190 (29.4)	198 (30.6)	161 (24.9)	167 (25.8)	114 (17.6)	183 (28.3)
GI-RADS 4	Probably malignant	40 (6.2)	46 (7.1)	171 (26.4)	78 (12.1)	43 (6.7)	32 (4.9)
GI-RADS 5	Very probably malignant	122 (18.8)	161 (24.9)	106 (16.4)	169 (26.1)	189 (29.2)	163 (25.2)
IOTA							
Benign	Only B features	402 (62.1)	501 (77.4)	471 (72.8)	411 (63.5)	494 (76.4)	451 (69.7)
Malignant	Only M features	162 (25.1)	68 (10.5)	117 (18.1)	137 (21.2)	79 (12.2)	161 (24.9)
Indeterminate	No features or both M and B features	83 (12.8)	78 (12.1)	59 (9.1)	99 (15.3)	74 (11.4)	35 (5.4)

Data are number of adnexal masses. Data in parentheses are percentages

O-RADS Ovarian-Adnexal Reporting and Data System, *GI-RADS* Gynecologic Imaging Reporting and Data System, *IOTA* International Ovarian Tumor Analysis, *B* benign, *M* malignant

the O-RADS categorized 441 (68.2%) as benign and 206 (31.8%) as malignant. The sensitivity, specificity, PPV, and NPV of O-RADS for the diagnosis of malignant AM were 96.6% (172/178), 92.8% (435/469), 83.5% (172/206), and 98.6% (435/441), respectively. Considering combined GI-RADS 4 and GI-RADS 5 as a predictor for malignancy, the GI-RADS categorized 452 (69.9%) as benign and 195 (30.1%) as malignant. The sensitivity, specificity, PPV, and NPV of GI-RADS for the diagnosis of malignant AM were 92.7% (165/178), 93.5% (439/469), 84.6% (165/195),

Table 3 Malignancy rates in the categories of three ultrasound classification systems

Category	Total no. (n = 647)	Final diagnosis		Recommended malignancy rate (%)	Calculated malignancy rate (%)	p value
		Benign (n = 469)	Malignant (n = 178)			
O-RADS						< 0.001
O-RADS 2	262 (40.5)	261 (55.7)	1 (0.6)	< 1	0.4	
O-RADS 3	179 (27.7)	174 (3.7.1)	5 (2.8)	1–< 10	2.8	
O-RADS 4	37 (5.7)	26 (5.5)	11 (6.2)	10–< 50	30.6	
O-RADS 5	169 (26.1)	8 (1.7)	161 (90.4)	≥ 50	95.3	
GI-RADS						< 0.001
GI-RADS 2	269 (41.6)	268 (57.1)	1 (0.6)	< 1	0.4	
GI-RADS 3	183 (28.3)	171 (36.5)	12 (6.7)	1–4	6.6	
GI-RADS 4	32 (4.9)	22 (4.7)	10 (5.6)	5–20	31.3	
GI-RADS 5	163 (25.2)	8 (1.7)	155 (87.1)	> 20	95.1	
IOTA						< 0.001
Benign	451 (69.7)	437 (93.2)	14 (7.9)	1–9	3.1	
Malignant	161 (24.9)	12 (2.6)	149 (83.7)	69–94	92.5	
Indeterminate	35 (5.4)	20 (4.2)	15 (8.4)	13–53	42.9	

Unless otherwise indicated, data are number of adnexal masses. Data in parentheses are percentages. The recommended malignancy rates are based on the literature

O-RADS Ovarian-Adnexal Reporting and Data System, *GI-RADS* Gynecologic Imaging Reporting and Data System, *IOTA* International Ovarian Tumor Analysis

Table 4 Diagnostic validity of three ultrasound classification systems using consensus data

	O-RADS	GI-RADS	IOTA
Cutoff	> O-RADS 3	> GI-RADS 3	Only M features, no features, or both M and B features
Sensitivity (%)	96.6 (172/178) [92.8–98.8]	92.7 (165/178) [87.83–96.1]	92.1 (164/178) [87.2–95.6]
Specificity (%)	92.8 (435/469) [90.0–94.9]	93.6 (439/469) [91–95.6]	93.2 (437/469) [90.5–95.3]
Positive predictive value (%)	83.5 (172/206) [77.7–88.3]	84.6 (165/195) [78.8–89.4]	83.7 (164/196) [77.7–88.6]
Negative predictive value (%)	98.6 (435/441) [97.1–99.5]	97.1 (437/450) [95.1–98.5]	96.9 (437/451) [94.9–98.3]
No. of true-positive findings	172	165	164
No. of false-negative findings	6	13	14
No. of false-positive findings	34	30	32
No. of true-negative findings	435	439	437

Data in parentheses were used to calculate percentages. Data in brackets are 95% confidence intervals

No. number, *O-RADS* Ovarian-Adnexal Reporting and Data System, *GI-RADS* Gynecologic Imaging Reporting and Data System, *IOTA* International Ovarian Tumor Analysis, *B* benign, *M* malignant

and 97.1% (439/452), respectively. Considering only M features, no features, or the presence of both M and B features as predictors for malignancy, the IOTA simple rules categorized 451 (69.7%) as benign and 196 (30.2%) as malignant. The sensitivity, specificity, PPV, and NPV of IOTA for the diagnosis of malignant AM were 92.1% (164/178), 93.2% (437/469), 83.7% (164/196), and 96.9% (437/451), respectively.

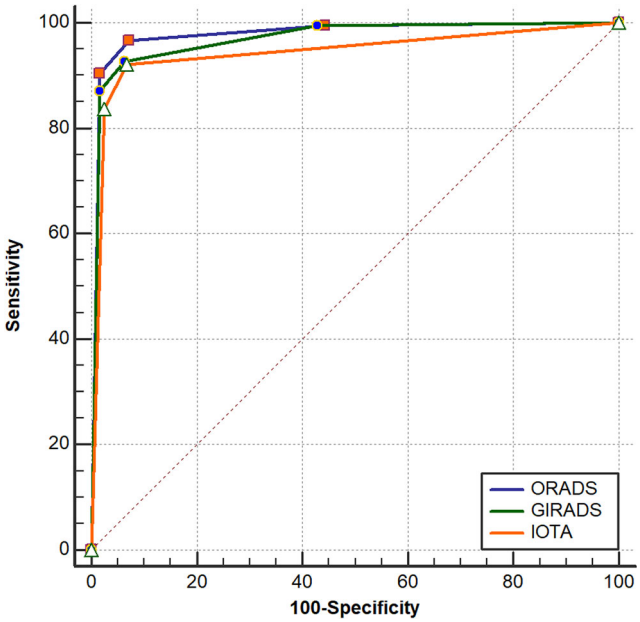

Fig. 2 Analysis of the ROC curves of the three ultrasound classification systems. ROC = receiver operating characteristic; O-RADS = Ovarian-Adnexal Reporting and Data System; GI-RADS = Gynecologic Imaging Reporting and Data System; IOTA = International Ovarian Tumor Analysis

Comparison of three US classification systems

In ROC analysis, the O-RADS had significantly higher area under the curve (AUC) than GI-RADS and IOTA (0.98, 95% CI = 0.96–0.99 vs 0.97, 95% CI = 0.95–0.98 and 0.94, 95% CI = 0.92–0.96; $p = 0.004$ and $p < 0.001$, respectively) (Fig. 2). The O-RADS had significantly higher sensitivity for malignancy than GI-RADS ($p = 0.003$) and IOTA ($p = 0.0007$), but non-significant slightly lower specificity ($p > 0.05$). No significant differences were detected between GI-RADS and IOTA regarding sensitivity and specificity ($p > 0.05$), but the GI-RADS had a significantly higher AUC than IOTA ($p < 0.001$).

Inter-reviewer agreement

IRA for O-RADS, GI-RADS, and IOTA stratified by the reviewer is listed in Table 5. O-RADS, GI-RADS, and IOTA showed similar overall IRA agreement ($\kappa = 0.77$, 0.69, and 0.63, respectively) with a tendency toward higher IRA with O-RADS than with GI-RADS and IOTA. The IRA for all reviewer combinations ranged between κ, 0.59, and 0.90 for O-RADS; between κ, 0.53, and 0.89 for GI-RADS; and between κ, 0.40, and 0.91 for IOTA.

Representative cases of our study are shown in Figs. 3, 4, 5, and 6.

Discussion

The comparison among the internationally established US classification systems of AM provides a systematic approach

Table 5 Inter-reviewer agreement of three ultrasound classification systems stratified by reviewer

Reviewer	System	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	All reviewers
Reviewer 1	O-RADS	0.79 (0.74–0.83)	0.70 (0.65–0.74)	0.59 (0.53–0.64)	0.78 (0.73–0.82)	
	GI-RADS	0.68 (0.63–0.73)	0.53 (0.47–0.58)	0.58 (0.52–0.63)	0.73 (0.68–0.77)	
	IOTA	0.45 (0.39–0.52)	0.65 (0.59–0.71)	0.91 (0.87–0.94)	0.40 (0.34–0.46)	
Reviewer 2	O-RADS		0.84 (0.80–0.87)	0.66 (0.61–0.71)	0.85 (0.81–0.88)	
	GI-RADS		0.66 (0.61–0.70)	0.89 (0.85–0.92)	0.74 (0.69–0.78)	
	IOTA		0.71 (0.65–0.77)	0.49 (0.62–0.56)	0.89 (0.85–0.93)	
Reviewer 3	O-RADS			0.81 (0.76–0.85)	0.90 (0.86–0.93)	
	GI-RADS			0.77 (0.72–0.81)	0.53 (0.48–0.58)	
	IOTA			0.75 (0.69–0.81)	0.66 (0.60–0.82)	
Reviewer 4	O-RADS				0.72 (0.71–0.76)	
	GI-RADS				0.74 (0.69–0.78)	
	IOTA				0.44 (0.38–0.50)	
All reviewers	O-RADS					0.77 (0.74–0.78)
	GI-RADS					0.69 (0.66–0.70)
	IOTA					0.63 (0.61–0.66)

Data are kappa values. Data in parentheses are 95% confidence intervals. The κ values were interpreted as follows: 0.00–0.20 = poor agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = good agreement, and 0.81–1.00 = very good agreement

O-RADS Ovarian-Adnexal Reporting and Data System, *GI-RADS* Gynecologic Imaging Reporting and Data System, *IOTA* International Ovarian Tumor Analysis

to patients with AM, helps determine the correct course of action, and facilitates the identification of malignant AM. In this study, we used three US classification systems—the O-RADS, GI-RADS, and IOTA simple rules—to classify AM. Although the risk stratification proposed by the GI-RADS and IOTA, including the recommendations for management and follow-up, has undergone successful prospective and external validation [6, 14, 16–24], to date, the performance of the O-RADS has not been tested. The present study was performed to determine the malignancy rates, validity, and reliability of the O-RADS when applied to a database of AM collected before the development of the system. To place the results in perspective, we also performed a similar analysis based on widely used recommendations from the GI-RADS and IOTA simple rules. Therefore, the present study compared the three established US classification systems.

The overall findings demonstrated that the three US classification systems had shown great value in the diagnosis of malignant AM; among them, the O-RADS performed the best. When considering combined O-RADS 4 and O-RADS 5 as a predictor for malignancy, the O-RADS had a statistically higher sensitivity for malignancy than GI-RAD and IOTA

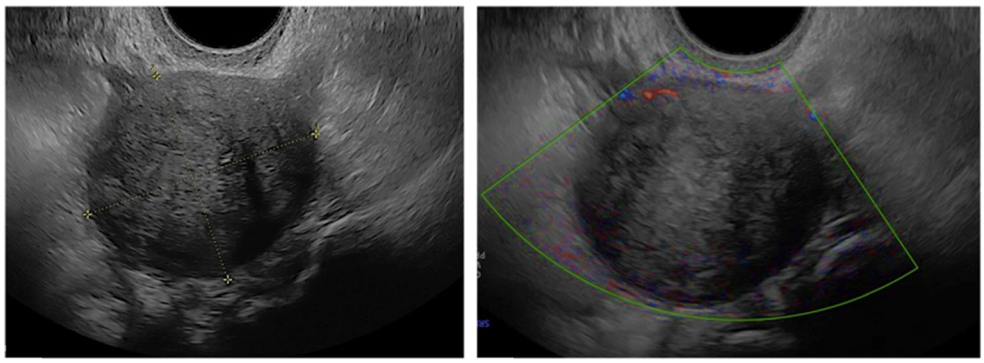

Fig. 3 A 41-year-old woman with a pathologically proven ovarian fibrothecoma. **a** Transvaginal gray-scale ultrasound reveals a 6-cm well-defined solid, smooth contours, heterogeneous echoic appearance mass containing acoustic shadowing in the right adnexa. **b** Color Doppler ultrasound reveals no flow (color score = 1). Based on consensus reviewing of the sonographic findings, the lesion was categorized as O-RADS 3, GI-RADS 4, and IOTA indeterminate. O-RADS = Ovarian-Adnexal Reporting and Data System; GI-RADS = Gynecologic Imaging Reporting and Data System; IOTA = International Ovarian Tumor Analysis

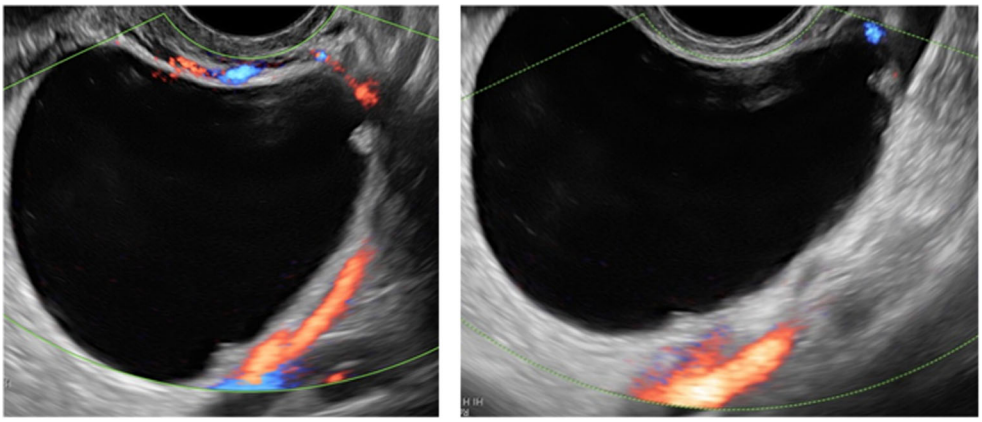

Fig. 4 A 48-year-old woman with a pathologically proven serous cystadenoma. **a** and **b** Transvaginal color Doppler ultrasound shows an 8-cm well-defined, thin-walled unilocular cyst in the right adnexa. The cyst has small solid components (two papillary projections of 6 mm maximal size) with no flow (color score = 1). Based on consensus reviewing of the sonographic findings, the lesion was categorized as O-RADS 4, GI-RADS 4, and IOTA benign. O-RADS = Ovarian-Adnexal Reporting and Data System; GI-RADS = Gynecologic Imaging Reporting and Data System; IOTA = International Ovarian Tumor Analysis

simple rules (p = 0.003 and 0.0007, respectively), but the specificity of three systems was quite comparable ($p > 0.05$). The positive predictive value (PPV) and negative predictive value (NPV) were quite high, as well as similar in three classification systems (PPV > 80% and NPV > 90%). The superior sensitivity of O-RADS can be mainly attributed to the comprehensive description and interpretation provided by the O-RADS for determining which AM requires no follow-up, conservative follow-up, or surgical removal. In contrast, the IOTA simple rules and GI-RADS did not provide adequate follow-up guidelines.

Notably, in our study, however, O-RADS demonstrated a non-significant slightly lower specificity (92.8%) in diagnosis of AM compared to GI-RADS (93.6%) and IOTA (93.2%), even though it did retain a specificity above 90% as per the consensus data. A decrease in specificity, however, indicates an increase in false-positive outcomes. In other terms, more benign AM could be diagnosed as malignant, and these patients may be referred to surgical procedures. Consequently, the low specificity of O-RADS can result in overtreatment of AM, which may be a major point of discussion. Although a false-positive diagnosis will not affect survival [13], this issue needs to be addressed on the further iteration of the O-RADS system.

From the ROC analysis, the performance of the O-RADS for the diagnosis of malignancy was higher than that of the GI-RADS and IOTA guidelines (p = 0.004 and < 0.001, respectively), because more malignant lesions were correctly classified into malignant categories with the O-RADS, resulting in an increase in true-positive results, as well as a decrease in false-negative results. This is also attributed to the appropriate management strategy assigned by the O-RADS, which is based on risk assessment with more conservative management for benign-appearing AM and referral to a gynecologic oncologist for suspicious AM.

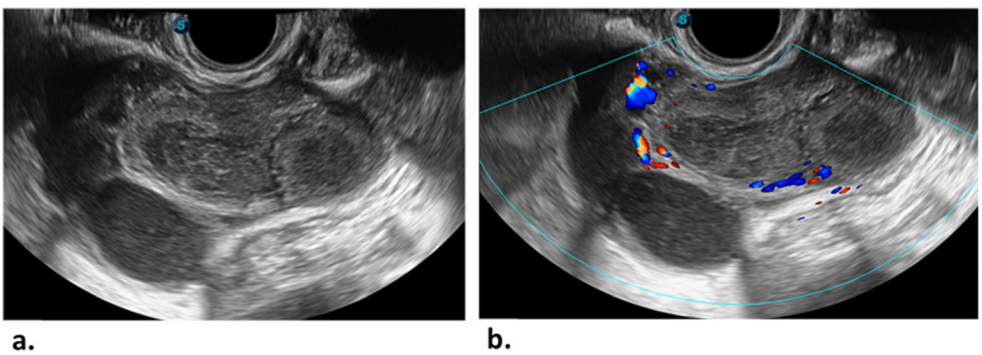

Fig. 5 A 35-year-old woman with a pathologically proved tubo-ovarian abscess. **a** Transvaginal gray-scale ultrasound demonstrates a 9-cm complex cystic lesion in the left adnexa. The lesion is composed of multilocular thick-walled folded cysts harboring turbid fluid content and internal reticulation. **b** Transvaginal color Doppler ultrasound reveals the color flow within the thick irregular walls of the abnormal tube; however, it is absent within cyst content. Based on consensus reviewing of the sonographic findings, the lesion was categorized as O-RADS 3, GI-RADS 4, and IOTA indeterminate. O-RADS = Ovarian-Adnexal Reporting and Data System; GI-RADS = Gynecologic Imaging Reporting and Data System; IOTA = International Ovarian Tumor Analysis

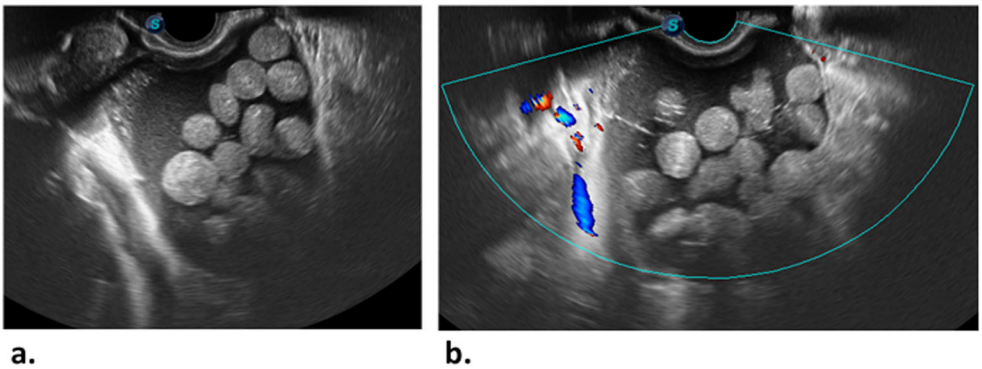

Fig. 6 A 27-year-old woman with a pathologically proven dermoid cyst. **a** Transvaginal gray-scale ultrasound demonstrates a 10-cm complex cystic lesion in the left adnexa. The cyst contains floating echogenic spherical structures. **b** Transvaginal color Doppler ultrasound reveals no flow (color score = 1). Based on consensus reviewing of the sonographic findings, the lesion was categorized as O-RADS 3, GI-RADS 3, and IOTA indeterminate. O-RADS = Ovarian-Adnexal Reporting and Data System; GI-RADS = Gynecologic Imaging Reporting and Data System; IOTA = International Ovarian Tumor Analysis

The basic examination to assess the malignancy rate of AM is the US. However, to date, few studies have discussed the recommended malignancy rate of AM by various US classification systems. We assessed the malignancy rates of AM in our study and found it increased with increasing suspicious sonographic patterns based on the O-RADS, GI-RADS, and IOTA categories. The malignancy rates were comparable to the recommended rates in O-RADS and IOTA categories [6, 13], whereas malignancy rates in GI-RADS categories were higher than the recommended rates [10].

One of the main disadvantages of the IOTA simple rules is that they yield an inconclusive result in about 24% of all AM [25]. Wherever the IOTA simple rules produce an inconclusive result, subjective assessment of US findings by an experienced US examiner is recommended, because this provides the most accurate diagnosis [22]. In our study, all US examinations were performed and reviewed by highly experienced radiologists, and the poor-quality and missing images were excluded. These factors, in addition to the consensus reviewing, may explain the lower prevalence of indeterminate AM in our study.

IRA and reproducibility are essential for assessing a classification system. Our study results show similar IRA in reporting AM when using O-RADS compared to GI-RADS and IOTA simple rules with a tendency toward increased IRA when O-RADS is used. IRA among all reviewers for O-RADS was 0.77; for GI-RADS, 0.69; and for IOTA, 0.63. The IRA in our study ranged from moderate to very good for three classification systems, indicating acceptable consistency, which is very similar to the previous studies [9, 14, 26, 27], which performed IRA for GI-RADS and IOTA.

All systems are expected to undergo periodic revisions. Some of the drawbacks described in the present study can be reduced. For instance, the low specificity of O-RADS can be increased by incorporating additional details in AM categorization and improving the recommendations for follow-up of AM that do not meet the criteria for malignancy. However, the new recommendations must balance the competing goals of reducing the number of missed malignant AM and minimizing the attention and resources committed to benign AM.

The strengths of the study are that the three classification systems were applied on a considerable multicenter database, which was valuable to demonstrate a minor but significant difference in the validity of the three systems. Additionally, all examinations were performed on the same patients, providing an ideal comparison in terms of patient characteristics. Furthermore, all radiologists who performed and reviewed the US examinations were highly experienced. However, our study has several limitations. First, we conducted our study retrospectively, and the analysis was based solely on static operator-dependent images instead of real practice, which may result in unavoidable bias. Second, all US examinations were performed by multiple radiologists and using different devices and transducers. Third, women with normal ovary (O-RADS 1 and GI-RADS 1), women with insufficient follow-up information, images with bad quality, and incomplete evaluation (O-RADS 0) were excluded from the study. This exclusion may result in selection bias, which could have led to a lower rate of benign AMs. So a large multicenter randomized controlled study is needed to avoid selection bias and validate our findings. Finally, these US-based classification systems can operate differently in different populations and practice conditions. Nevertheless, we believe that these limitations would similarly affect the three systems and would not negate the comparative results that we obtained.

Conclusion

In conclusion, the O-RADS compares favorably with GI-RADS and IOTA simple rules. The O-RADS had higher sensitivity than GI-RADS and IOTA simple rules with relatively

similar specificity and reliability. The O-RADS can become a widely respected classification system for US structured reporting of AM. However, a further increase of specificity, reduction in complexity, and refined imaging characteristics may be required. These results, along with future prospective longitudinal studies, should help guide revisions of all US classification systems and perhaps lead to a single unified multidisciplinary system that can be accepted internationally.

Acknowledgments The authors thank all staff members and colleagues in the Radiology and Obstetrics& Gynecology Departments at Zagazig and Benha Universities, and Al-Ahrar Teaching Hospital for their helpful cooperation.

Funding information The authors state that this work has not received any funding.

Availability of data and materials The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is the corresponding author: Mohammad Abd Alkhalik Basha.

Conflict of interest The author of this manuscript declare no relevant conflicts of interest, and no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry The corresponding author has great statistical expertise

Informed consent Written informed consent was obtained from all patients.

Ethical approval Institutional review boards' approval was obtained.

Study subjects or cohorts overlap A subset of the data, on 258 of the 609 women, was previously published [Basha, M.A.A., Refaat, R., Ibrahim, S.A. et al. Eur Radiol (2019) 29: 5981. https://doi.org/10.1007/s00330-019-06181-0]. The prior study evaluated diagnostic performance and IRA of the GI-RADS for diagnosis of AM by US; in the current study, we compared the O-RADS with GI-RADS and IOTA regarding their validity and reliability for the US diagnosis of AM.

Methodology
- Retrospective
- Diagnostic or prognostic study
- Performed at multiple centres

References

1. American College of Radiology. ACR appropriateness criteria 2008: clinically suspected adnexal mass. American College of Radiology Web site. Available at http://www.acr.org/SecondaryMainMenuCategories/quality_safety/app_criteria/pdf/ExpertPanelonWomensImaging/SuspectedAdnexalMassesDoc11.aspx. Accessed 9 Nov 2009
2. Liu J, Xu Y, Wang J (2007) Ultrasonography, computed tomography and magnetic resonance imaging for diagnosis of ovarian carcinoma. Eur J Radiol 62:328–334
3. Levine D, Brown DL, Andreotti RF et al (2010) Management of asymptomatic ovarian and other adnexal cysts imaged at US: Society of Radiologists in Ultrasound consensus conference statement. Radiology 256:943–954
4. Timmerman D, Valentin L, Bourne TH et al (2000) 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 16:500–505
5. Timmerman D, Testa AC, Bourne T et al (2008) Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol 31:681–690
6. Timmerman D, Van Calster B, Testa A et al (2016) Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group. Am J Obstet Gynecol 214:424–437
7. Van Calster B, Van Hoorde K, Valentin L et al (2014) 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 349:g5920
8. Van Calster B, Timmerman D, Valentin L et al (2012) Triaging women with ovarian masses for surgery: observational diagnostic study to compare RCOG guidelines with an International Ovarian Tumour Analysis (IOTA) group protocol. BJOG 119:662–671
9. Amor F, Vaccaro H, Alcázar JL, León M, Craig JM, Martinez J (2009) Gynecologic imaging reporting and data system: a new proposal for classifying adnexal masses on the basis of sonographic findings. J Ultrasound Med 28:285–291
10. Amor F, Alcázar JL, Vaccaro H, León M, Iturra A (2011) GI-RADS reporting system for ultrasound evaluation of adnexal masses in clinical practice: a prospective multicenter study. Ultrasound Obstet Gynecol 38:450–455
11. Kaijser J, Sayasneh A, Van Hoorde K et al (2014) Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. Hum Reprod Update 20:449–462
12. Andreotti RF, Timmerman D, Benacerraf BR et al (2018) Ovarian-adnexal reporting lexicon for ultrasound: a white paper of the ACR Ovarian-Adnexal Reporting and Data System Committee. J Am Coll Radiol 15:1415–1429
13. Andreotti RF, Timmerman D, Strachowski LM et al (2019) O-RADS US Risk Stratification and Management System: a consensus guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. Radiology 294:168–185
14. Basha MAA, Refaat R, Ibrahim SA et al (2019) Gynecology Imaging Reporting and Data System (GI-RADS): diagnostic performance and inter-reviewer agreement. Eur Radiol 29:5981–5990
15. Heintz AP, Odicino F, Maisonneuve P et al (2003) Carcinoma of the ovary. Int J Gynecol Obstet 83:135–166
16. Orozco Fernández R, Peces Rama A, Llanos Llanos MC, Martinez Mendoza A, Machado Linde F, Nieto Diaz A (2015) Clinical application of the gynecologic imaging reporting and data system (GI-RADS) for the evaluation of adnexal masses. SM J Gynecol Obstet 1:1009–1012
17. Zhang T, Li F, Liu J, Zhang S (2017) Diagnostic performance of the Gynecology Imaging Reporting and Data System for malignant adnexal masses. Int J Gynaecol Obstet 137:325–331
18. Migda M, Bartosz M, Migda MS, Kierszk M, Katarzyna G, Maleńczyk M (2018) Diagnostic value of the gynecology imaging reporting and data system (GI-RADS) with the ovarian malignancy marker CA-125 in preoperative adnexal tumor assessment. J Ovarian Res 11:92

19. Rams N, Muñoz R, Soler C, Parra J (2015) Resultados de la clasificación Gynecologic Imaging Reporting and Data System para la catalogación de masas anexiales. Prog Obstet Ginecol 58:125–129
20. Alcazar JL, Pascual MA, Graupera B et al (2016) External validation of IOTA simple descriptors and simple rules for classifying adnexal masses. Ultrasound Obstet Gynecol 48:397–402
21. Koneczny J, Czekierdowski A, Florczak M, Poziemski P, Stachowicz N, Borowski D (2017) The use of sonographic subjective tumor assessment, IOTA logistic regression model 1, IOTA Simple Rules and GI-RADS system in the preoperative prediction of malignancy in women with adnexal masses. Ginekol Pol 88:647–653
22. Timmerman D, Ameye L, Fischerova D et al (2010) Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. BMJ 341:c6839
23. Garg S, Kaur A, Mohi JK, Sibia PK, Kaur N (2017) Evaluation of IOTA simple ultrasound rules to distinguish benign and malignant ovarian tumours. J Clin Diagn Res 11:TC06
24. Wynants L, Timmerman D, Verbakel JY et al (2017) 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 23:5082–5090
25. Timmerman D, Testa AC, Bourne T et al (2005) 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 23:8794–8801
26. Zannoni L, Savelli L, Jokubkiene L et al (2006) Intra-and inter-observer agreement with regard to describing adnexal masses using International Ovarian Tumor Analysis (IOTA) terminology: a reproducibility study involving seven observers. Ultrasound Obstet Gynecol 44:100 –108
27. Ruiz de Gauna B, Sanchez P, Pineda L, Utrilla-Layna J, Juez L, Alcazar JL (2014) Interobserver agreement in describing adnexal masses using the International Ovarian Tumor Analysis simple rules in a real-time setting and using three-dimensional ultrasound volumes and digital clips. Ultrasound Obstet Gynecol 44:95–99