World Journal of Radiology

Submit a Manuscript: https://www.f6publishing.com

World J Radiol 2022 September 28; 14(9): 319-328

DOI: 10.4329/wjr.v14.i9.319

ISSN 1949-8470 (online)

ORIGINAL ARTICLE

Retrospective Study

Reliability of ultrasound ovarian-adnexal reporting and data system amongst less experienced readers before and after training

Prayash Katlariwala, Mitchell P Wilson, Yeli Pi, Baljot S Chahal, Roger Croutze, Deelan Patel, Vimal Patel, Gavin Low

Specialty type: Radiology, nuclear medicine and medical imaging

Provenance and peer review: Unsolicited article; Externally peer reviewed.

Peer-review model: Single blind

Peer-review report's scientific quality classification

Grade A (Excellent): 0
Grade B (Very good): B
Grade C (Good): C, C
Grade D (Fair): 0
Grade E (Poor): 0

P-Reviewer: Aydin S, Turkey; Sahin H, Turkey

Received: March 12, 2022
Peer-review started: March 12, 2022
First decision: May 31, 2022
Revised: June 14, 2022
Accepted: September 13, 2022
Article in press: September 13, 2022
Published online: September 28, 2022

Prayash Katlariwala, Mitchell P Wilson, Yeli Pi, Baljot S Chahal, Roger Croutze, Deelan Patel, Vimal Patel, Gavin Low, Department of Radiology and Diagnostic Imaging, University of Alberta, Edmonton T6G 2B7, AB, Canada

Corresponding author: Prayash Katlariwala, BSc, Department of Radiology and Diagnostic Imaging, University of Alberta, 8440-112 Street NW, Edmonton T6G 2B7, AB, Canada. prayashvk@gmail.com

Abstract

BACKGROUND

The 2018 ovarian-adnexal reporting and data system (O-RADS) guidelines are aimed at providing a system for consistent reports and risk stratification for ovarian lesions found on ultrasound. It provides key characteristics and findings for lesions, a lexicon of descriptors to communicate findings, and risk characterization and associated follow-up recommendation guidelines. However, the O-RADS guidelines have not been validated in North American institutions or amongst less experienced readers.

AIM

To evaluate the diagnostic accuracy and inter-reader reliability of ultrasound O-RADS risk stratification amongst less experienced readers in a North American institution with and without pre-test training.

METHODS

A single-center retrospective study was performed using 100 ovarian/adnexal lesions of varying O-RADS scores. Of these cases, 50 were allotted to a training cohort and 50 to a testing cohort *via* a non-randomized group selection process in order to approximately equal distribution of O-RADS categories both within and between groups. Reference standard O-RADS scores were established through consensus of three fellowship-trained body imaging radiologists. Three PGY-4 residents were independently evaluated for diagnostic accuracy and inter-reader reliability with and without pre-test O-RADS training. Sensitivity, specificity, positive predictive value, negative predictive value (NPV), and area under the curve (AUC) were used to measure accuracy. Fleiss kappa and weighted quadratic (pairwise) kappa values were used to measure inter-reader reliability. Statistical significance was $P < 0.05$.

RESULTS
Mean patient age was 40 ± 16 years with lesions ranging from 1.2 to 22.5 cm. Readers demonstrated excellent specificities (85%-100% pre-training and 91%-100% post-training) and NPVs (89%-100% pre-training and 91-100% post-training) across the O-RADS categories. Sensitivities were variable (55%-100% pre-training and 64%-100% post-training) with malignant O-RADS 4 and 5 Lesions pre-training and post-training AUC values of 0.87-0.95 and 0.94-098, respectively ($P < 0.001$). Nineteen of 22 (86%) misclassified cases in pre-training were related to mischaracterization of dermoid features or wall/septation morphology. Fifteen of 17 (88%) of post-training misclassified cases were related to one of these two errors. Fleiss kappa inter-reader reliability was 'good' and pairwise inter-reader reliability was 'very good' with pre-training and post-training assessment (k = 0.76 and 0.77; and k = 0.77-0.87 and 0.85-0.89, respectively).

CONCLUSION
Less experienced readers in North America achieved excellent specificities and AUC values with very good pairwise inter-reader reliability. They may be subject to misclassification of potentially malignant lesions, and specific training around dermoid features and smooth *vs* irregular inner wall/septation morphology may improve sensitivity.

Key Words: Ovarian-adnexal reporting and data system; Ovary; Malignancy; Accuracy; Reliability; Ultrasound

Core Tip: This study supports the applied utilization of the ovarian-adnexal reporting and data system (O-RADS) ultrasound risk stratification tool by less experienced readers in North America. KEY RESULTS: The O-RADS ultrasound risk stratification requires validation in less experienced North American readers; Excellent specificities (85%-100%), area under the curve values (0.87-0.98) and very good pairwise reliability can be achieved by trainees in North America regardless of formal pre-test training; Less experienced readers may be subject to down-grade misclassification of potentially malignant lesions and specific training about typical dermoid features and smooth *vs* irregular margins of ovarian lesions may help improve sensitivity.

Citation: Katlariwala P, Wilson MP, Pi Y, Chahal BS, Croutze R, Patel D, Patel V, Low G. Reliability of ultrasound ovarian-adnexal reporting and data system amongst less experienced readers before and after training. *World J Radiol* 2022; 14(9): 319-328
URL: https://www.wjgnet.com/1949-8470/full/v14/i9/319.htm
DOI: https://dx.doi.org/10.4329/wjr.v14.i9.319

INTRODUCTION

Building on the original ovarian-adnexal reporting and data system (O-RADS) publication in 2018, the American College of Radiology (ACR) O-RADS working group has recently introduced risk stratification and management recommendations to supplement the detailed reporting lexicon for this classification system[1,2]. These guidelines aim to provide consistent language, accurate characterization, and standardized recommendations for ovarian/adnexal lesions identified on ultrasound, ultimately improving the quality of communication between ultrasound examiners, referring clinicians and patients. A couple of recent papers have validated the use of the O-RADS system as an effective tool for the detection of ovarian malignancies, possessing high diagnostic accuracy and robust inter-reader reliability even without formalized training[3,4] For its future directions, the O-RADS working group specifically calls for additional studies validating this system in North American institutions and amongst less experienced readers[1]. Thus, the primary objective of the present study is to assess the inter-reader reliability of O-RADS classification amongst North American Radiology trainees using the O-RADS system, before and after training.

MATERIALS AND METHODS

This is a single center retrospective study performed at the University of Alberta Institutional Health

Research Ethics Board (HREB) approval was acquired prior to the study (Pro00097690). Patient consent for individual test cases was waived by the HREB as cases were retrospectively retrieved from the institutional Picture Archiving and Communication System (PACS) and de-identified prior to review by individual readers.

Patient selection

The University of Alberta institutional PACS was reviewed between May 2017 and July 2020 for all pelvic ultrasounds in adult female patients that demonstrated at least 1 ovarian/adnexal lesion with adequate diagnostic quality, including the presence of transvaginal 2D and Doppler sonographic image of the lesion(s) of interest. Studies were excluded if limited by technical factors such as bowel gas, large size of lesion, location of the adnexa, or inability to tolerate transvaginal ultrasound (O-RADS 0)[1].

A total of 100 diagnostic non-consecutive cases were selected by a Steering Committee of three authors including the senior author (Wilson MP, Patel V, Low G). In patients with more than one ovarian lesion, only different ipsilateral lesions were used with each individual lesion extracted as an independent blinded case when presented to study readers and the lesion of interest was designated with an arrow in each respective case. No concurrent contralateral lesions were used within the same patient. Cases were selected non-consecutively to acquire an approximately equal range of O-RADS 1 to O-RADS 5 Lesions. From these 100 cases, 50 cases were selected into separate 'Training' and 'Testing' groups. All cases were then de-identified leaving only the age, with 50 years of age used as a threshold for menopausal status. The cases were then listed as a teaching file in our institutional PACS (IMPAX 6 AGFA Healthcare) with a randomly assigned case number. All available static and cine imaging for the case were included in the teaching case file, with the additional inclusion of a 'key image' identifying the lesion intended for risk stratification with an arrow.

Training and testing

Three PGY-4 Diagnostic Radiology residents from a single institution volunteered as readers for the present study, henceforth referred to as R1, R2 and R3. The residents did not have prior formal experience with the O-RADS, SRU or IOTA systems for adnexal lesions, but have been exposed to ultrasonography in routine clinical practice totaling up to 12 wk. The residents were provided a copy of the O-RADS US Risk Stratification and Management System publication for independent review[1], and subsequently were asked to independently analyze all 50 'Testing' cases assigning the best O-RADS risk stratification score and lexicon descriptor. Answers were collected using an online Google Forms survey. Following completion of the testing file, an interval of six weeks was selected to prevent case recall. The senior author (Low G) then provided residents with a presentation reviewing the O-RADS system including lexicon descriptors, differentiating nuances for scoring, and separate examples of lesions in each O-RADS category (no overlap with cases used in the study design). The residents were then provided access to the 50 'Training' cases together with an answer key, for practice purposes and to establish familiarity with using the O-RADS system. Following the training session, and after the readers had reviewed the 'Training Cases,' the 50 "Testing" cases were then re-randomized, and independently scored again by all 3 readers in similar fashion to the pre-training format.

For both pre and post-training assessment, the reference gold standard was determined by independent consensus reading of three fellowship-trained body imaging radiologists with experience in gynaecologic ultrasound with 5, 13, and > 25 years of ultrasound experience (Wilson MP, Patel V, Low G).

Statistical analysis

The diagnostic accuracy of each individual reader and inter-observer variability between each reader both pre-training and post-training was evaluated. Continuous variables were expressed as the mean ± standard deviation. Statistical tests included: Fleiss kappa (overall agreement) and weighted quadratic kappa (pairwise agreement) was used to calculate the inter-reader agreement. The kappa (k) value interpretation as suggested by Cohen was used: κ < 0.20 (poor agreement), κ = 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)[5]. Diagnostic accuracy measurements including sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated per O-RADS category for each individual reader. Receiver operating characteristic (ROC) analysis was used to evaluate the area under the receiver operating curve (AUC) for each reader. All statistical analyses were conducted using IBM SPSS (version 26) and MedCalc (version 19.6.1). A P value of < 0.05 was considered as statistically significant.

RESULTS

Cumulatively, the testing portion of the study was comprised of 50 cases. The average age of the patients in the test cohort was 40.1 ± 16.2 years and a range from 17 to 85 years. According to the reference standard, there were 10 cases (20%) of O-RADS 1, 10 cases (20%) of O-RADS 2, 7 cases (14%) of

Table 1 Sensitivity, specificity, positive predictive value and negative predictive value per ovarian-adnexal reporting and data system category for each reader on the pre-training assessment

Pre training	ORADS 1, %	ORADS 2, %	ORADS 3, %	ORADS 4, %	ORADS 5, %
Sensitivity					
R1	90 (55.5 to 99.8)	100 (69.5 to 100)	100 (59.0 to 100)	92 (61.5 to 99.8)	55 (23.4 to 83.3)
R2	90% (55.5 to 99.8)	100% (69.2 to 100)	71 (29.0 to 96.3)	92 (61.5 to 99.8)	82 (48.2 to 97.7)
R3	90 (55.5 to 99.8)	100 (69.2 to 100)	100 (59.0 to 100)	75 (42.8 to 94.5)	55 (23.4 to 83.3)
Specificity					
R1	100 (91.2 to 100)	85 (70.2 to 94.3)	98 (87.7 to 99.4)	100 (90.8 to 100)	100 (91.0 to 100)
R2	100 (91.2 to 100)	90 (76.3 to 97.2)	98 (87.7 to 99.4)	97 (86.2 to 99.9)	100 (91.0 to 100)
R3	98 (86.8 to 99.9)	90 (76.3 to 97.2)	95 (84.2 to 99.4)	95 (82.3 to 99.4)	100 (91.0 to 100)
PPV					
R1	100	63 (44.4 to 77.7)	88 (50.2 to 98.0)	100	100
R2	100	71 (49.7 to 86.4)	83 (40.5 to 97.4)	92 (61.2 to 98.7)	100
R3	90 (56.2 to 98.4)	71 (49.7 to 86.4)	78 (47.5 to 93.1)	82 (52.9 to 94.8)	100
NPV					
R1	98 (86.2 to 99.6)	100	100	97 (85.3 to 99.6)	89 (80.3 to 93.7)
R2	98 (86.2 to 99.6)	100	96 (86.7 to 98.6)	97 (85.0 to 99.6)	95 (84.8 to 98.6)
R3	98 (85.9 to 99.6)	100	100	93 (81.8 to 97.0)	89 (80.3 to 93.7)

O-RADS: Ovarian-Adnexal Reporting and Data System; PPV: Positive predictive value; NPV: Negative predictive value.

O-RADS 3, 12 cases (24%) of O-RADS 4 and 11 cases (22%) of O-RADS 5. Of the complete test cohort, 24 lesions (48%) were lateralized to the left and right with 2 lesions (4%) being located centrally in the pelvis and with an indeterminate origin site.

Overall, the lesion sizes ranged from 1.2 cm to 22.5 cm with an average size of 6.9 ± 4.7. Mean lesion size by O-RADS category was: 2.1 ± 0.5 cm for O-RADS 1, 5.1 ± 1.4 cm for O-RADS 2, 10.6 ± 5.8 cm for O-RADS 3, 7.8 ± 4.6 cm for O-RADS 4 and 9.4 ± 4.4 cm for O-RADS 5 ($P < 0.001$).

Inter-reader reliability

The overall inter-reader agreement for the 3 readers as a group on the pre-training assessment was considered '*good*' (k = 0.76 [0.68 to 0.84, 95% Confidence Interval {CI}], p < 0.001). Kappa values for agreement on individual 0-RADS categories were '*good*' or '*very good*', as follows: O-RADS 1, k = 0.82 (0.66 to 0.98), $P < 0.001$; O-RADS 2, k = 0.78 (0.62 to 0.94), $P < 0.001$; O-RADS 3, k = 0.74 (0.58 to 0.90), $P < 0.001$; O-RADS 4, k = 0.73 (0.57 to 0.89), $P < 0.001$; O-RADS 5, k = 0.72 (0.56 to 0.88), $P < 0.001$.

The overall inter-reader agreement for the 3 readers as a group on the post-training assessment was considered '*good*' (k = 0.77 [0.69 to 0.86, 95%CI], $P < 0.001$). Kappa values for agreement on individual O-RADS categories were '*good*' or '*very good*', as follows: O-RADS 1, k = 0.96 (0.80 to 1), $P < 0.001$; O-RADS 2, k = 0.81 (0.65 to 0.97), $P < 0.001$; O-RADS 3, k = 0.65 (0.49 to 0.81), $P < 0.001$; O-RADS 4, k = 0.74 (0.58 to 0.90), $P < 0.001$; O-RADS 5, k = 0.70 (0.54 to 0.86), $P < 0.001$.

Pairwise inter-reader agreement, as evaluated using weighted kappa, was '*very good*', as follows: Pre-training: R1 and R2, k = 0.79 (0.62 to 0.96), $P < 0.001$; R1 and R3, k = 0.77 (0.59 to 0.95) $P < 0.001$; R2 and R3, k = 0.87 (0.73 to 1.00) $P < 0.001$. Post-training: R1 and R2, k = 0.86 (0.73 to 0.99), $P < 0.001$; R1 and R3, k = 0.85 (0.71 to 0.99) $P < 0.001$; R2 and R3, k = 0.89 (0.78 to 0.99) $P < 0.001$.

Diagnostic accuracy

The respective sensitivity, specificity, NPV, and PPV for each reader per O-RADS category are included in Table 1 for the pre-training assessment and Table 2 for the post-training assessment. All readers showed excellent specificities (85%-100% pre-training and 91%-100% post-training) and NPVs (89%-100% pre-training and 91%-100% post-training) across the O-RADS categories. Sensitivities range from 90%-100% in both pre-training and post-training for O-RADS 1 and O-RADS 2, 71%-100% pre-training and 86%-100% post-training for O-RADS 3, 75-92% in both pre-training and post-training for O-RADS 4, and 55%-82% pre-training and 64%-82% post-training for O-RADS 5. Readers misclassified 22 (14.7%) of 150 cases on pre-training assessment and 17 (11.3%) on post-training assessment. Misclassified cases and their respective lexicon descriptors are included in Table 3.

Table 2 The sensitivity, specificity, positive predictive value and negative predictive value per Ovarian-Adnexal Reporting and Data System category for each reader on the post-training assessment

Post training	ORADS 1, %	ORADS 2, %	ORADS 3, %	ORADS 4, %	ORADS 5, %
Sensitivity					
R1	100 (69.2 to 100)	100 (69.2 to 100)	100 (59 to 100)	92 (61.5 to 99.8)	73 (39 to 94)
R2	90 (55.5 to 99.8)	90 (55.5 to 99.8)	86 (42.1 to 99.6)	92 (61.5 to 99.8)	82 (48.2 to 97.7)
R3	100 (69.2 to 100)	100 (69.2 to 100)	100 (59 to 100)	75 (42.8 to 94.5)	64 (30.8 to 89.1)
Specificity					
R1	100 (91.2 to 100)	95 (83.1 to 99.4)	98 (87.7 to 99.9)	97 (86.2 to 99.9)	100 (91 to 100)
R2	100 (91.2 to 100)	98 (86.8 to 99.9)	93 (80.9 to 98.5)	95 (82.3 to 99.4)	100 (91 to 100)
R3	100 (91.2 to 100)	95 (83.1 to 99.4)	91 (77.9 to 97.4)	97 (86.2 to 99.9)	100 (91 to 100)
PPV					
R1	100	83 (56.4 to 95.1)	88 (50.2 to 98)	92 (61.2 to 98.7)	100
R2	100	90 (56.2 to 98.4)	67 (39.2 to 86.1)	85 (58.5 to 95.5)	100
R3	100	83 (56.4 to 95.1)	64 (40.8 to 81.7)	90 (55.9 to 98.5)	100
NPV					
R1	100	100	100	97 (85 to 99.6)	93 (83.2 to 97.2)
R2	98 (86.2 to 99.6)	98 (85.9 to 99.6)	98 (86.7 to 99.6)	97 (84.6 to 99.6)	95 (84.8 to 98.6)
R3	100	100	100	93 (82.2 to 97.1)	91 (81.7 to 95.5)

O-RADS: Ovarian-adnexal reporting and data system; PPV: Positive predictive value; NPV: Negative predictive value.

The ROC analysis evaluated diagnostic accuracy of the readers are included in Figure 1A for the pre-training assessment and Figure 1B for the post-training assessment. Given that higher O-RADS score (*i.e.* O-RADS 4 and O-RADS 5) are predictors of malignancy, reader AUC values are as follows: Pre-training: R1, AUC of 0.87 (0.75 to 0.95), $P < 0.001$; R2, AUC of 0.95 (0.84 to 0.99), $P < 0.001$; R3, AUC of 0.89 (0.77 to 0.96), $P < 0.001$. Post-training: R1, AUC of 0.96 (0.86 to 0.99), $P < 0.001$; R2, AUC of 0.98 (0.89 to 1.00), $P < 0.001$; R3, AUC of 0.94 (0.83 to 0.99), $P < 0.001$.

Pairwise comparison of the ROC curves showed a significant improvement post-training *vs* pre-training for R1 ($P = 0.04$) but not for R2 ($P = 0.29$) and R3 ($P = 0.21$).

DISCUSSION

This study demonstrates 'good' to 'very good' inter-reader agreement amongst less experienced readers in a North American institution, with pairwise and overall kappa values between spanning 0.76 and 0.89 ($P < 0.001$). The high degree of reliability is concordant with the findings of a prior study by Cao *et al*[4]. In their study performed at a tertiary care hospital and a cancer hospital in China, the pair-wise inter-reader agreement between a first-year radiology resident and a staff radiologist with 9 years experience in gynaecologic ultrasound was assessed. The authors found a kappa of 0.714 for the O-RADS system and a kappa of 0.77 for classifying lesion categories ($P < 0.001$).

Our study also highlights excellent diagnostic accuracies of resident readers when compared to a reference standard of three body-fellowship trained radiologists with experience in gynaecologic ultrasound. Solely with self-review of the O-RADS guidelines, the readers achieved high specificities greater than 0.85 and NPV greater than 0.89. These results persisted post-training, showing significant improvement in 1 resident ($P = 0.04$) and a trend towards improved accuracy amongst the other readers. The otherwise non-significant differences are due in part to excellent overall diagnostic accuracy without pre-test training as well as inadequate power to detect small differences. The study suggests that individual review of the O-RADS risk stratification is sufficient in less experienced readers with respect to specificity and AUC values. In this regard, this study validates the use of O-RADS risk classification amongst less experienced readers in a North American institution; a cohort specifically requiring validation by the ACR O-RADS committee[1].

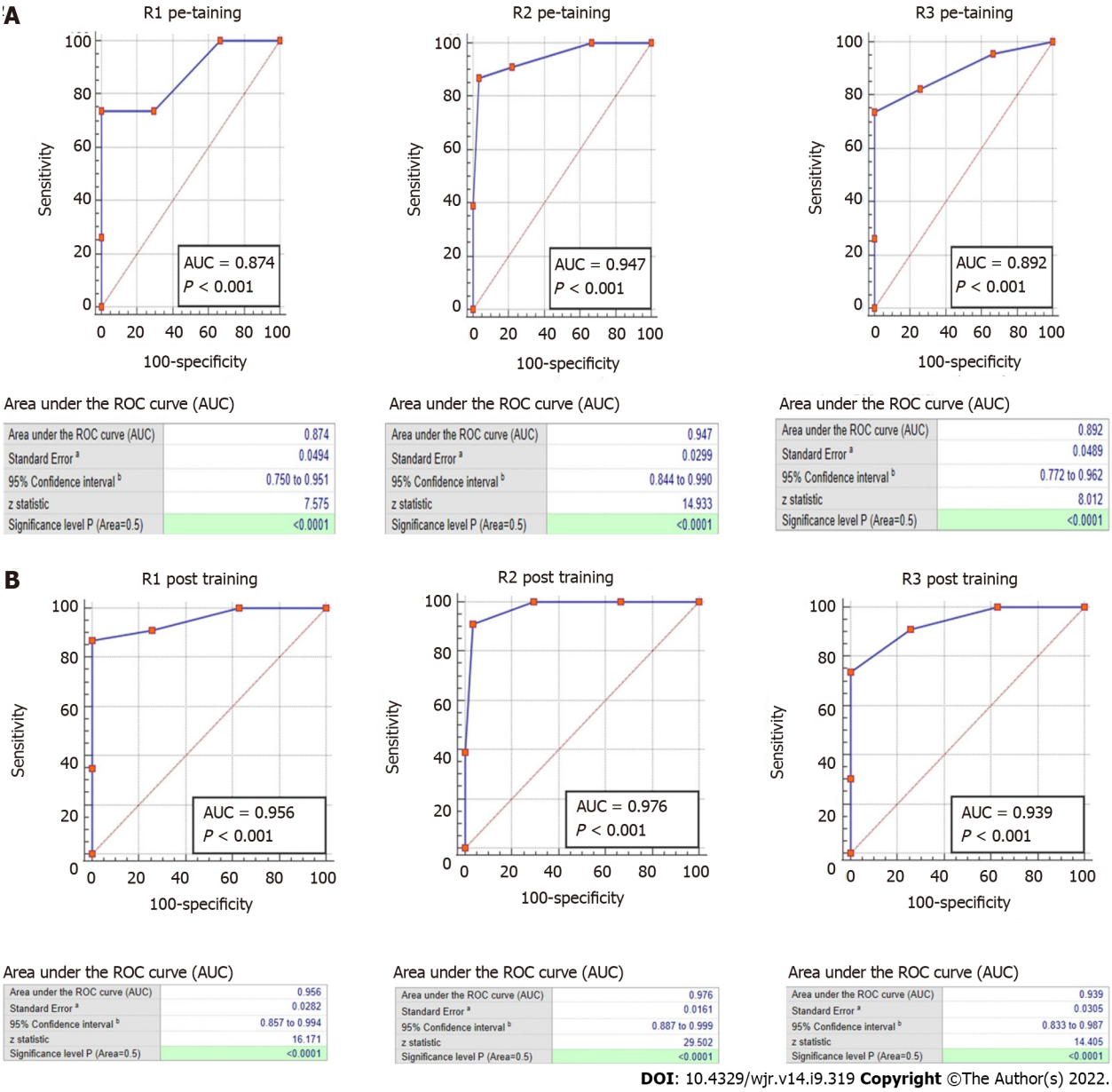
An important risk amongst less experienced readers is the potential to misclassify potentially malignant lesions as benign. The sensitivity results in this study were variable in both pre-training and post-training assessment, particularly in higher O-RADS categories. In their respective pre-training and

Table 3 Misclassified ovarian-adnexal reporting and data system categories by readers in pre-training and post-training assessment

ORADS category	Reference standard lexicon descriptor	Misclassification category	Reader lexicon descriptor	Frequency of error in pre-training	Frequency of error in post-training
ORADS 1	Follicle defined as a simple cyst ≤ 3 cm	ORADS 2	Follicle defined as a simple cyst ≤ 3 cm	1	1
	Follicle defined as a simple cyst ≤ 3 cm	ORADS 2	Simple cyst > 5 cm but < 10 cm	1	0
	Follicle defined as a simple cyst ≤ 3 cm	ORADS 3	Multilocular cyst with smooth inner walls/septations < 10 cm, CS1-3	1	0
ORADS 2	simple cyst > 3 cm to 5 cm	ORADS 3	Unilocular cyst with irregular inner wall < 3mm height, any size	0	1
ORADS 3	Multilocular cyst with smooth inner walls/septations, < 10 cm, CS1-3	ORADS 2	Simple cyst > 5 cm but < 10 cm	1	0
	Multilocular cyst with smooth inner walls/septations, < 10 cm, CS1-3	ORADS 4	Multilocular cyst, irregular inner wall ± irregular septation	0	1
	Unilocular cyst (simple or non-simple) ≥ 10 cm	ORADS 4	Unilocular cyst with 1-3 papillary projections	1	0
ORADS 4	Multilocular cyst, irregular inner wall ± irregular septation	ORADS 1	Follicle defined as a simple cyst ≤ 3 cm	1	0
	Multilocular cyst, irregular inner wall ± irregular septation	ORADS 2	Classic benign lesion (hemorrhagic cyst < 10 cm)	1	0
	Multilocular cyst, irregular inner wall ± irregular septation	ORADS 3	Typical dermoid cyst, endometrioma, hemorrhagic cyst ≥ 10 cm	0	1
	Multilocular cyst, irregular inner wall ± irregular septation	ORADS 3	Multilocular cyst with smooth inner walls/septations < 10 cm, CS1-3	3	4
ORADS 5	Solid lesion with irregular outer contour	ORADS 2	Classic benign lesion (dermoid cyst < 10 cm)	10	4
	Solid lesion with irregular outer contour	ORADS 3	Solid lesion with smooth outer contour, any size, CS = 1	0	1
	Solid lesion with irregular outer contour	ORADS 3	Typical dermoid cyst, endometrioma, hemorrhagic cyst ≥ 10 cm	0	1
	Solid lesion with irregular outer contour	ORADS 4	Unilocular cyst with solid component	1	1
	Solid lesion with irregular outer contour	ORADS 4	Solid lesion with smooth outer contour, any size, CS = 2-3	0	2
	Multilocular cyst with solid component, CS3-4	ORADS 4	Multilocular cyst with solid component, CS1-2	1	0

O-RADS: Ovarian-adnexal reporting and data system; CS: Color scor.

post-training assessments, sensitivities were 64%-82% and 75%-92% for O-RADS 4 and 55%-82% and 64%-82% for O-RADS 5. The most frequent error on pre-training assessment was classifying a solid lesion as O-RADS 2 with a "typical dermoid cyst < 10 cm" lexicon descriptor. This error accounted for 45% (10/22) of misclassified cases in the pre-training assessment, with a reduction to 27% (4/17) of misclassified cases following training. This pitfall may be mitigated by comparing the hyperechoic component of a solid ovarian lesion to the surrounding pelvic and subcutaneous fat. The lesion should be classified as a dermoid only if it is isoechoic to the internal reference, and/or demonstrates one of three typical features including: (1) hyperechoic component with shadowing; (2) hyperechoic lines and dots; or (3) floating echogenic spherical structures[1,2]. In reviewing the test cases, all the solid lesions misclassified as dermoid had echogenicity lower than the intrapelvic fat. An example of this misclassification is shown in Figure 2.

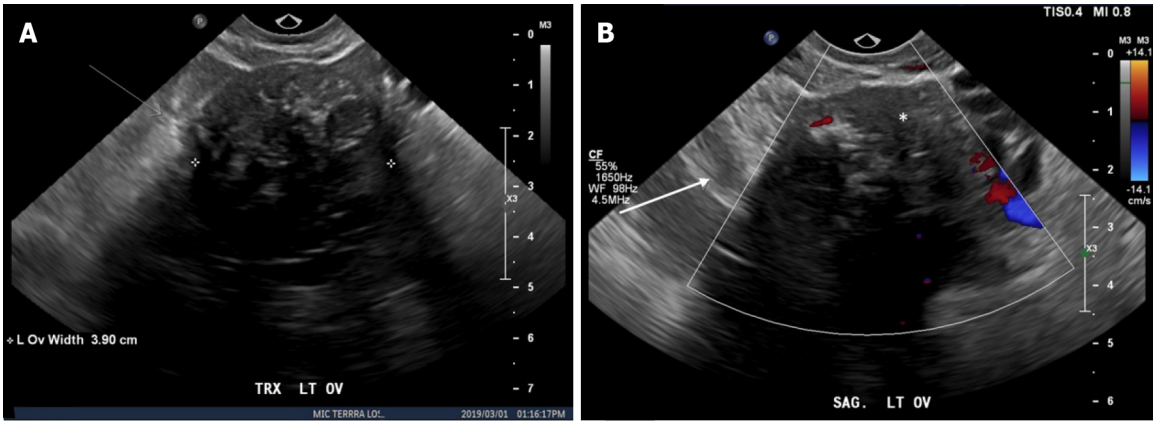


DOI: 10.4329/wjr.v14.i9.319

Figure 1 Receiver operating characteristic curve. A: Receiver operating characteristic (ROC) curve of each reader on the pre-training assessment; B: ROC curve of each reader on the post-training assessment. AUC: Area under the curve.

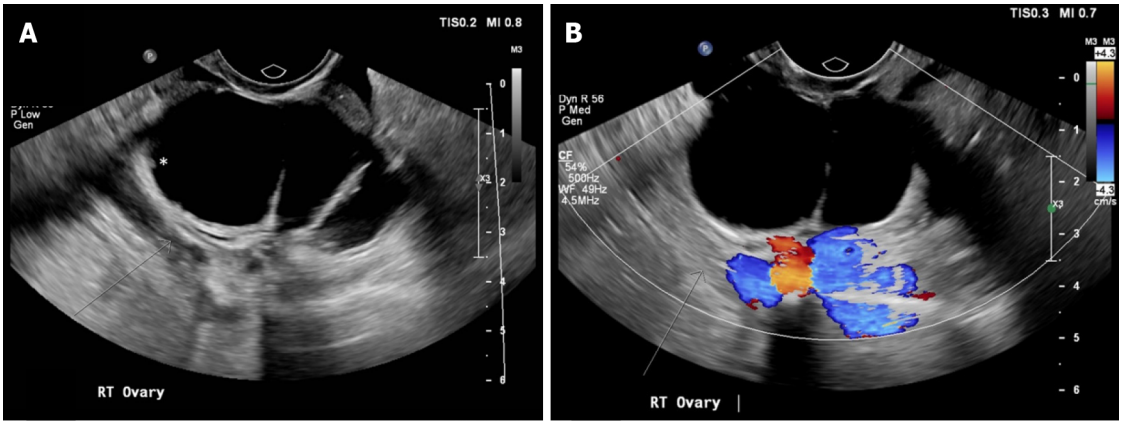
A second frequent error occurred in multilocular lesions with an irregular inner wall and/or irregular septation (O-RADS 4). These lesions were downgraded to O-RADS 1 through O-RADS 3 Lesions with variable lexicon descriptors used. Most commonly, these were characterized as a multilocular lesion with a smooth inner wall (O-RADS 3) in both pre-training and post-training assessment, suggesting that specific training on this finding was not sufficient in the current study. In this scenario, it is important that readers comprehensively evaluate the entire lesion on the cine clips, as irregularity in the inner wall/septation may be a subtle finding only seen in a small area within the lesion. An example of this misclassification is shown in Figure 3. Unlike the dermoid misclassification, however, this downgrade still results in a recommendation for evaluation by an ultrasound specialist or MRI and gynecology referral, reducing the risk for adverse potential complication of this misclassification. Despite these misclassifications, the negative predictive value in O-RADS 4 and O-RADS 5 Lesions remains high in both pre-training and post-training assessment (89%-97% and 91%-97%).

This study is subject to several limitations Firstly, this was a retrospective non-consecutive review. As the menopausal status was often not provided in the clinical information, an arbitrary age cut-off of 50 years was used to differentiate pre-menopausal (< 50 years) *vs* post-menopausal patients (≥ 50 years), an approach has also been used in previous epidemiologic studies[6-8]. Secondly, we did not use a pathological reference standard. Our reference standard was an expert panel of 3 three fellowship-trained radiologists with experience in gynaecologic ultrasound. However, as O-RADS is a risk stratification system that is designed to be applied universally in the clinical setting and as our study is



DOI: 10.4329/wjr.v14.i9.319 **Copyright** ©The Author(s) 2022.

Figure 2 An example of a left ovarian solid lesion misclassified as a typical ovarian dermoid. A: Static gray-scale images; B: Static color Doppler ultrasound images. Static gray-scale and color Doppler ultrasound images shows a solid hypoechoic lesion with a non-uniform (irregular) margin demonstrated on the color Doppler image (Ovarian-Adnexal Reporting and Data System 5). The lesion demonstrates punctate echogenic areas (white asterisk) which are less echogenic than the surrounding pelvic fat (white arrow). Further, the echogenic areas do not fulfill one of the three descriptors required to characterize as a "typical dermoid cyst < 10 cm" according to ovarian-adnexal reporting and data system criteria (2). The hypoechoic lesion with posterior shadowing suggests a fibrous lesion.



DOI: 10.4329/wjr.v14.i9.319 **Copyright** ©The Author(s) 2022.

Figure 3 An example of a right ovarian cystic lesion misclassified as a "multilocular cyst < 10 cm, smooth inner wall, color score 1-3" (Ovarian-Adnexal Reporting and Data System 3). A: Static gray-scale images; B: Static color Doppler ultrasound images. Static gray-scale and color Doppler ultrasound images show a multilocular cyst with a subtle non-uniform (irregular) inner wall with solid components < 3 mm in height (white asterisk) (ovarian-adnexal reporting and data system 4) (2).

designed primarily to evaluate inter-reader agreement, an expert consensus panel is arguably a reasonable reference standard, and one that simulates 'real world' clinical practice. A similar approach has been taken in previous O-RADS accuracy studies[3,9]. Thirdly, our sample size of 50 training cases was fairly small. A large multi-center inter-observer variability study in North America would be useful to evaluate the generalizability of our findings. Despite these limitations, we believe that the rigorous study design and specific reader cohort provide valuable insight into a needed area of validation identified by the ACR O-RADS committee.

CONCLUSION

In summary, the study validated the use of the ACR-ORADS risk stratification system in less experienced readers, showing excellent specificities and AUC values when compared to a consensus reference standard and high pairwise inter-reader reliability. Less experienced readers may be at risk for misclassification of potentially malignant lesions, and specific training around common pitfalls may help improve sensitivity.

ARTICLE HIGHLIGHTS

Research background

The 2018 Ovarian-Adnexal Reporting and Data System (O-RADS) guidelines are aimed at providing a system for consistent reports and risk stratification for ovarian lesions found on ultrasound. It provides key characteristics and findings for lesions, a lexicon of descriptors to communicate findings, and risk characterization and associated follow-up recommendation guidelines. However, the O-RADS guidelines have not been validated in North American institutions.

Research motivation

The O-RADS ultrasound risk stratification requires validation in less experienced North American readers.

Research objectives

Evaluate the diagnostic accuracy and inter-reader reliability of ultrasound O-RADS risk stratification amongst less experienced readers in a North American institution without and with pre-test training.

Research methods

A single-center retrospective study was performed using 100 ovarian/adnexal lesions of varying O-RADS scores. Of these cases, 50 were allotted to a training cohort and 50 to a testing cohort *via* a non-randomized group selection process in order to approximately equal distribution of O-RADS categories both within and between groups. Reference standard O-RADS scores were established through consensus of three fellowship-trained body imaging radiologists. Three PGY-4 residents were independently evaluated for diagnostic accuracy and inter-reader reliability without and with pre-test O-RADS training. Sensitivity, specificity, positive predictive value, negative predictive value, and area under the curve (AUC) were used to measure accuracy. Fleiss kappa and weighted quadratic (pairwise) kappa values were used to measure inter-reader reliability.

Research results

Excellent specificities (85%-100%), AUC values (0.87-0.98) and very good pairwise reliability can be achieved by trainees in North America regardless of formal pre-test training. Less experienced readers may be subject to down-grade misclassification of potentially malignant lesions and specific training about typical dermoid features and smooth *vs* irregular margins of ovarian lesions may help improve sensitivity.

Research conclusions

Less experienced readers in North America achieved excellent specificities and AUC values with very good pairwise inter-reader reliability though they may be subject to misclassification of potentially malignant lesions. Training around dermoid features and smooth *vs* irregular inner wall/septation morphology may improve sensitivity.

Research perspectives

This study supports the applied utilization of the O-RADS ultrasound risk stratification tool by less experienced readers in North America.

FOOTNOTES

Author contributions: All authors contributed equally to the paper.

Supported by RSNA Research & Education Foundation Medical Student Grant #RMS2020.

Institutional review board statement: Institutional Health Research Ethics Board (HREB) approval was acquired from the University of Alberta prior to the study (Pro00097690).

Informed consent statement: Institutional ethics approval was obtained for this study which also waived the requirement for the informed consent. Please see institutional HREB approval document for details.

Conflict-of-interest statement: All authors have no conflicts of interest.

Data sharing statement: No additional data available.

Country/Territory of origin: Canada

ORCID number: Prayash Katlariwala 0000-0002-5822-1071; Mitchell P Wilson 0000-0002-1630-5138; Vimal Patel 0000-0003-2972-5980; Gavin Low 0000-0002-4959-8934.

S-Editor: Liu JH
L-Editor: A
P-Editor: Liu JH

REFERENCES

1 **Andreotti RF**, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL, Bourne T, Brown DL, Coleman BG, Frates MC, Goldstein SR, Hamper UM, Horrow MM, Hernanz-Schulman M, Reinhold C, Rose SL, Whitcomb BP, Wolfman WL, Glanc P. O-RADS US Risk Stratification and Management System: A Consensus Guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. *Radiology* 2020; **294**: 168-185 [PMID: 31687921 DOI: 10.1148/radiol.2019191150]

2 **Andreotti RF**, Timmerman D, Benacerraf BR, Bennett GL, Bourne T, Brown DL, Coleman BG, Frates MC, Froyman W, Goldstein SR, Hamper UM, Horrow MM, Hernanz-Schulman M, Reinhold C, Strachowski LM, Glanc P. Ovarian-Adnexal Reporting Lexicon for Ultrasound: A White Paper of the ACR Ovarian-Adnexal Reporting and Data System Committee. *J Am Coll Radiol* 2018; **15**: 1415-1429 [PMID: 30149950 DOI: 10.1016/j.jacr.2018.07.004]

3 **Pi Y**, Wilson MP, Katlariwala P, Sam M, Ackerman T, Paskar L, Patel V, Low G. Diagnostic accuracy and inter-observer reliability of the O-RADS scoring system among staff radiologists in a North American academic clinical setting. *Abdom Radiol (NY)* 2021; **46**: 4967-4973 [PMID: 34185128 DOI: 10.1007/s00261-021-03193-7]

4 **Cao L**, Wei M, Liu Y, Fu J, Zhang H, Huang J, Pei X, Zhou J. Validation of American College of Radiology Ovarian-Adnexal Reporting and Data System Ultrasound (O-RADS US): Analysis on 1054 adnexal masses. *Gynecol Oncol* 2021; **162**: 107-112 [PMID: 33966893 DOI: 10.1016/j.ygyno.2021.04.031]

5 **Landis JR**, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159-174 [PMID: 843571]

6 **Phipps AI**, Ichikawa L, Bowles EJ, Carney PA, Kerlikowske K, Miglioretti DL, Buist DS. Defining menopausal status in epidemiologic studies: A comparison of multiple approaches and their effects on breast cancer rates. *Maturitas* 2010; **67**: 60-66 [PMID: 20494530 DOI: 10.1016/j.maturitas.2010.04.015]

7 **Hill K**. The demography of menopause. *Maturitas* 1996; **23**: 113-127 [PMID: 8735350 DOI: 10.1016/0378-5122(95)00968-x]

8 **Im SS**, Gordon AN, Buttin BM, Leath CA 3rd, Gostout BS, Shah C, Hatch KD, Wang J, Berman ML. Validation of referral guidelines for women with pelvic masses. *Obstet Gynecol* 2005; **105**: 35-41 [PMID: 15625139 DOI: 10.1097/01.AOG.0000149159.69560.ef]

9 **Basha MAA**, Metwally MI, Gamil SA, Khater HM, Aly SA, El Sammak AA, Zaitoun MMA, Khattab EM, Azmy TM, Alayouty NA, Mohey N, Almassry HN, Yousef HY, Ibrahim SA, Mohamed EA, Mohamed AEM, Afifi AHM, Harb OA, Algazzar HY. Comparison of O-RADS, GI-RADS, and IOTA simple rules regarding malignancy rate, validity, and reliability for diagnosis of adnexal masses. *Eur Radiol* 2021; **31**: 674-684 [PMID: 32809166 DOI: 10.1007/s00330-020-07143-7]

Published by **Baishideng Publishing Group Inc**
7041 Koll Center Parkway, Suite 160, Pleasanton, CA 94566, USA
Telephone: +1-925-3991568
E-mail: bpgoffice@wjgnet.com
Help Desk: https://www.f6publishing.com/helpdesk
https://www.wjgnet.com