76110_Auto_Edited.docx

WORD COUNT	3323	TIME SUBMITTED	23-AUG-2022 03:52PM
		PAPER ID	89778590

Name of Journal: *World Journal of Radiology*
Manuscript NO: 76110
Manuscript Type: ORIGINAL ARTICLE

Retrospective Study

Inter-reader reliability of ultrasound O-RADS risk stratification amongst less experienced readers in a North American institution before and after training

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

Abstract

BACKGROUND

The 2018 O-RADS guideline 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 for to communicate findings, and risk characterization and associated follow-up recommendation guideline. However, the O-RADS guideline has not been validated in North American institutes or amongst less experienced readers.

AIM

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.

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 (PPV), 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.

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 ***. Institutional Health Research Ethics Board (HREB) approval was acquired prior to the study (Pro***). 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 *** 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 (**, **, **). 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 totalling 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 (**) 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 (**, **, **).

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: $\kappa < 0.20$ (poor agreement), $\kappa = 0.21–0.40$ (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 O-RADS 3, 12 cases (24%) of O-RADS 4 and 11 cases (22%) of O-RADS 5. Off 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**.

The ROC analysis evaluated diagnostic accuracy of the readers are included in **Figure 1** for the pre-training assessment and **Figure 2** 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 concordance 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 US 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 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 3**.

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 4**. 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 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 O-RADS guideline 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 for to communicate findings, and risk characterization and associated follow-up recommendation guideline. However, the O-RADS guideline has not been validated in North American institutes.

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 (PPV), 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.

Research results

There is 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 experiences 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.

76110_Auto_Edited.docx

ORIGINALITY REPORT

4%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.pubfacts.com Internet	62 words — 2%
2	Xiao Zhu, Bo Peng, QiFeng Yi, Jia Liu, Jin Yan. "Prediction Model of Immunosuppressive Medication Non-adherence for Renal Transplant Patients Based on Machine Learning Technology", Frontiers in Medicine, 2022 Crossref	18 words — < 1%
3	Se Jin Lee, Hye Rim Oh, Sunghun Na, Han Sung Hwang, Seung Mi Lee. "Ultrasonographic ovarian mass scoring system for predicting malignancy in pregnant women with ovarian mass", Obstetrics & Gynecology Science, 2021 Crossref	15 words — < 1%
4	Yang Du, Meredith Bara, Prayash Katlariwala, Roger Croutze et al. "Effect of training on resident inter-reader agreement with American College of Radiology Thyroid Imaging Reporting and Data System", World Journal of Radiology, 2022 Crossref	15 words — < 1%
5	"Clinical Radiology Orals", Journal of Medical Imaging and Radiation Oncology, 2021 Crossref	12 words — < 1%

6 Marten E. van den Berg, Bertine M.J. Flokstra-de Blok, Berber J. Vlieg-Boerstra, Marjan Kerkhof et al. "Parental Eczema Increases the Risk of Double-Blind, Placebo-Controlled Reactions to Milk but Not to Egg, Peanut or Hazelnut", International Archives of Allergy and Immunology, 2012 — 12 words — < 1%
Crossref

7 Yang Du, Meredith Bara, Prayash Katlariwala, Roger Croutze et al. "Effect of training on resident inter-reader agreement with American College of Radiology Thyroid Imaging Reporting and Data System", World Journal of Radiology — 12 words — < 1%
Internet

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS