Diagn Interv Radiol 2024; DOI: 10.4274/dir.2024.242784

ABDOMINAL IMAGING

ORIGINAL ARTICLE

Diagnostic performance of the O-RADS MRI system for magnetic resonance imaging in discriminating benign and malignant adnexal lesions: a systematic review, meta-analysis, and meta-regression

Gülsüm Kılıçkap

Ankara Bilkent City Hospital, Clinic of Radiology, Ankara, Türkiye

PURPOSE

After the introduction of the Ovarian-Adnexal Reporting and Data System (O-RADS) for magnetic resonance imaging (MRI), several studies with diverse characteristics have been published to assess its diagnostic performance. This systematic review and meta-analysis aimed to assess the diagnostic performance of O-RADS MRI scoring for adnexal masses, accounting for the risk of selection bias.

METHODS

The PubMed, Scopus, Web of Science, and Cochrane databases were searched for eligible studies. Borderline or malignant lesions were considered malignant. All O-RADS MRI scores ≥4 were considered positive. The quality of the studies was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool. The pooled sensitivity, specificity, and likelihood ratio (LR) values were calculated, considering the risk of selection bias.

RESULTS

Fifteen eligible studies were found, and five of them had a high risk of selection bias. Between-study heterogeneity was low-to-moderate for sensitivity but substantial for specificity (I^2 values were 35.5% and 64.7%, respectively). The pooled sensitivity was significantly lower in the studies with a low risk of bias compared with those with a high risk of bias (93.0% and 97.5%, respectively; $P = 0.043$), whereas the pooled specificity was not different (90.4% for the overall population). The negative and positive LRs were 0.08 [95% confidence interval (CI) 0.05–0.11] and 10.0 (95% CI 7.7–12.9), respectively, for the studies with low risk of bias and 0.03 (95% CI 0.01–0.10) and 10.3 (95% CI 3.8–28.3), respectively, for those with high risk of bias.

CONCLUSION

The overall diagnostic performance of the O-RADS system is very high, particularly for ruling out borderline/malignant lesions, but with a moderate ruling-in potential. Studies with a high risk of selection bias lead to an overestimation of sensitivity.

CLINICAL SIGNIFICANCE

The O-RADS system demonstrates considerable diagnostic performance, particularly in ruling out borderline or malignant lesions, and should routinely be used in practice. The high between-study heterogeneity observed for specificity suggests the need for improvement in the consistent characterization of the benign lesions to reduce false positive rates.

KEYWORDS

Adnexal mass, ovarian cancer, Ovarian-Adnexal Reporting and Data System, magnetic resonance imaging, meta-analysis, systematic review

Corresponding author: Gülsüm Kılıçkap

E-mail: gkilickap@yahoo.com.tr

Received 28 March 2024; revision requested 12 May 2024; accepted 29 May 2024.

Epub: 08.07.2024

Publication date: xx.xx.2024

DOI: 10.4274/dir.2024.242784

You may cite this article as: Kılıçkap G. Diagnostic performance of the O-RADS MRI system for magnetic resonance imaging in discriminating benign and malignant adnexal lesions: a systematic review, meta-analysis, and meta-regression. *Diagn Interv Radiol.* 08 July 2024 DOI: 10.4274/dir.2024.242784 [Epub Ahead of Print].

Ovarian cancer is one of the leading causes of cancer-related death in women. Accurate characterization of adnexal masses is crucial for correct diagnosis and the prevention of unnecessary surgery. Transvaginal ultrasound is the first-line diagnostic method due to its relative affordability and widespread availability. However, magnetic resonance imaging (MRI) offers several advantages over ultrasound, including better characterization and visualization of the origin of the mass and higher resolution. Recently, the American College of Radiology (ACR) proposed a method-the Ovarian-Adnexal Reporting and Data System (O-RADS) for MRI (O-RADS MRI)- to standardize the analysis of adnexal masses.[1]

Following the introduction of the O-RADS MRI system, several studies assessing its validity, including a small number of meta-analyses, have been published.[2-18] These studies have diverse characteristics and were conducted at single or multicenter sites with varying levels of expertise and patient volumes. Given the increasing number of studies on the diagnostic value of the O-RADS MRI score in recent years and the potential heterogeneity among them, this study aims to conduct an updated systematic review and meta-analysis of these studies by taking into consideration their risk of bias. Therefore, this systematic review, meta-analysis, and meta-regression aim to assess the diagnostic value of the O-RADS MRI system in assessing adnexal masses and to reveal the rule-in and rule-out potential of borderline or malignant adnexal masses. Unlike other meta-analyses, the objective is to calculate the pooled sensitivity and specificity of O-RADS according to whether the studies included in the analysis are at high or low risk of patient selection bias.

Main points

- The diagnostic performance of the Ovarian-Adnexal Reporting and Data System (O-RADS) for magnetic resonance imaging (MRI) system is very high.
- The O-RADS MRI system is valuable in ruling out borderline or malignant adnexal masses.
- The ruling-in potential of the O-RADS system is moderate.
- Studies with a high risk of bias lead to overestimation of the sensitivity.

Methods

This systematic review and meta-analysis were prepared and presented in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) recommendations.[19] Since the data were obtained from manuscripts, informed consent was not required, and ethics committee approval was waived.

Study population and research question

The study population and research question were structured according to the PICO format (P- population, I- intervention/index test, C- comparator/reference test, and O- outcome) and included patients who underwent pelvic MRI examinations for adnexal masses. Studies were excluded if any of the following criteria were present: 1) absence of the standard reference test, 2) O-RADS scoring using non-MRI methods, 3) case-control studies or inappropriate selection or exclusion, 4) studies in which only specific lesions (such as only cystic lesions) or a specific O-RADS category were evaluated, and 5) studies assessing O-RADS scoring with non-contrast MRI, as this is not included in the standards proposed by the original O-RADS MRI scoring.

The index test was based on O-RADS MRI scoring, in which a score ≥4 was considered positive, and its diagnostic value was compared with the reference standard test result.

The comparison was made using the pathology or reasonable follow-up as a reference test. Borderline or malignant lesions were considered malignant.

The outcomes were diagnostic performance measures that included sensitivity, specificity, summary receiver operating characteristics (SROC) curve, and likelihood ratios (LRs).

Searching and extracting the data

The PubMed, Scopus, Web of Science, and Cochrane Central Register of Controlled Trials databases were searched for eligible studies on December 29, 2023. The search terms used in the PubMed database were as follows: "("Ovarian"[Title/Abstract] OR "adnexal"[Title/Abstract] OR "pelvic"[Title/Abstract]) AND ("Cancer"[Title/Abstract] OR "malignan*"[Title/Abstract] OR "tumor"[Title/Abstract] OR "mass*"[Title/Abstract] OR "lesion"[Title/Abstract]) AND ("O-RADS"[Title/Abstract] OR "ORADS"[Title/Abstract] OR "Ovarian adnexal reporting and data system"[Title/Abstract]) AND ("magnetic resonance imaging"[MeSH Terms] OR ("Magnetic Resonance"[Title/Abstract] OR "MRI"[Title/Abstract] OR "MR"[Title/Abstract]))." The same search terms were used in other databases with slight modifications to conform to the database's rules. No restriction (including language) was applied to the database searches.

The selection of the eligible studies and the number of manuscripts obtained from each database are shown in the PRISMA flowchart (Figure 1). After removing duplicated manuscripts, the titles and abstracts were initially screened for eligible studies, followed by a subsequent screening of the full-text manuscripts. One of the eligible studies was published in Chinese, and the full-text manuscript could not be obtained.[6] However, the abstract contained the required information to conduct a diagnostic meta-analysis; therefore, no eligible studies were discarded in the analysis.

For studies in which more than one investigator evaluated the MRI scores, the measurements of the most experienced investigator were used. If the most experienced investigator made more than one measurement, the first measurement was included in the analysis.

Lesion-based O-RADS MRI scoring was analyzed. Since lesion-based data could not be obtained in one study,[16] data for patient-based assessments given in the article were included in the analysis.

Assessment of the quality of the included studies

The quality of each eligible study was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool.[20] This tool includes four domains (patient selection, index test, reference standard, and flow and timing) to evaluate the risk of bias and applicability of primary diagnostic accuracy studies. Each study was scored for both risk of bias and concern for applicability as high, unclear, or low. Critical appraisal of the selected studies was conducted by two reviewers independently, and any discrepancies were resolved through consensus.

Certainty of evidence

Certainty of evidence was assessed using the Grading of Recommendations, Assessment, Development, and Evaluations tool.[21] As the pooled sensitivity was significantly different for the studies with low and high risk of bias, certainty of evidence was provided for sensitivity for the studies with low risk of bias. However, as there was no significant

difference in specificities between the studies with low and high risk of bias, certainty of evidence for specificity was given for the overall group.

Statistical analysis

Using the cut-off value of O-RADS scores ≥4, the number of the true positive, false positive, true negative, and false negative results were recorded, and sensitivity and specificity values were calculated. The data were pooled using bivariate random effects model meta-analysis and presented as a forest plot and the SROC curve. Random effects meta-regression analysis was performed by including the variable of patient selection bias, which was obtained with the QUADAS-2 tool (Model 1). In the case of significant relative sensitivity or specificity for the selection bias categories of high-risk versus low-risk group, the corresponding diagnostic measure was presented separately. Then, the age and the proportion of borderline or malignant cases were included in the meta-regression (Model 2). The mean (or median) age was not presented in the two studies;[6,12] therefore, these missing values were replaced with the overall mean obtained from the remaining studies. The performances of Model 1 and Model 2 were compared using the LR test.

The bivariate random effects model uses an unstructured variance-covariance matrix as the default method. The model was also run with the independent variance–covariance matrix to test whether the simpler (parsimonious) model is appropriate. Then the two models with different matrix structures were compared using the Akaike information criteria (AIC). As the model with an unstructured variance-covariance matrix had a lower (better) value of AIC, it is presented here.

The pooled estimates for positive and negative LR for O-RADS MRI scoring in diagnosing borderline or malignant lesions were calculated. It is generally accepted that a positive LR of >10 and negative LR of <0.1 are valuable in confirming or excluding the disease, respectively, while values of 5–10 and 0.1–0.2, respectively, are moderately effective in this regard.[22] The point estimates and their 95% confidence interval (CI) for positive and negative LR values (the LR matrix) were plotted to visually assess the confirming or excluding potential of O-RADS MRI scoring. Additionally, Fagan's nomogram was plotted to calculate the post-test probability of having borderline or malignant lesions. The mean and median values of the borderline/malignancy lesion proportions were 25.5% and 24.4%, respectively. Therefore, Fagan's nomogram was plotted using the pre-test probability value of 25% for borderline/malignant lesions.

The between-study heterogeneity was assessed using the I^2 statistics proposed by Zhou and Dendukuri[23] and also with Cochran's Q statistics and its *P* value. The I^2 parameter has values of 0%–100%; the values >50% and >75% are considered moderate and severe heterogeneity, respectively. Publication bias was assessed with a funnel plot proposed by Deeks et al.[24] and tested statistically. A *P value* of <0.05 was considered significant. Statistical analyses were performed using Stata version 17 (StataCorp, TX, USA), and the "metadata" and "midas" packages were used for the analysis.

Results

Fifteen eligible studies were found. The PRISMA flowchart for the selection of these studies is provided in Figure 1, and the characteristics of these studies are given in Table 1. Five of the studies were considered to have a high risk of patient selection bias based on the QUADAS-2 report. Figure 2 summarizes the interpretation with the QUADAS report, and the details are given in Supplementary Table 1.

The mean age ranged from 35.9 to 57 years, with a mean and standard deviation of 46.1 ± 7.1 years and a median and interquartile range (IQR) of 48.7 (40.0–50.8) years. The median proportion of borderline or malignant lesions was 25.2% (IQR 13.5%–38.8%). For the studies with a low risk of selection bias, this ranged from 11.2% to 52.9%, with a mean ± standard deviation of 25.5 ± 13.3 years and a median and IQR of 24.4 (13.5–31.4) years; for those with a high risk of bias, the range was from 11.8% to 65.4%, with a mean ± standard deviation of 32.4 ± 22.1 years and a median and IQR of 28.3 (14.3–42.0) years.

Meta-analysis of the eligible studies

The sensitivity values ranged from 81% to 100%, while specificity values ranged from 58.0% to 97.9%. In the analysis stratified for the risk of selection bias, there was low-to-moderate between-study heterogeneity for diagnostic sensitivity [I^2 values were 35.5% for the overall group and 39.8% and 14.2% for the studies with low risk and high risk of selection bias, respectively. The corresponding Cochran's Q statistics (*P* values) were 21.71 (*P* = 0.085), 14.95 (*P* = 0.092),

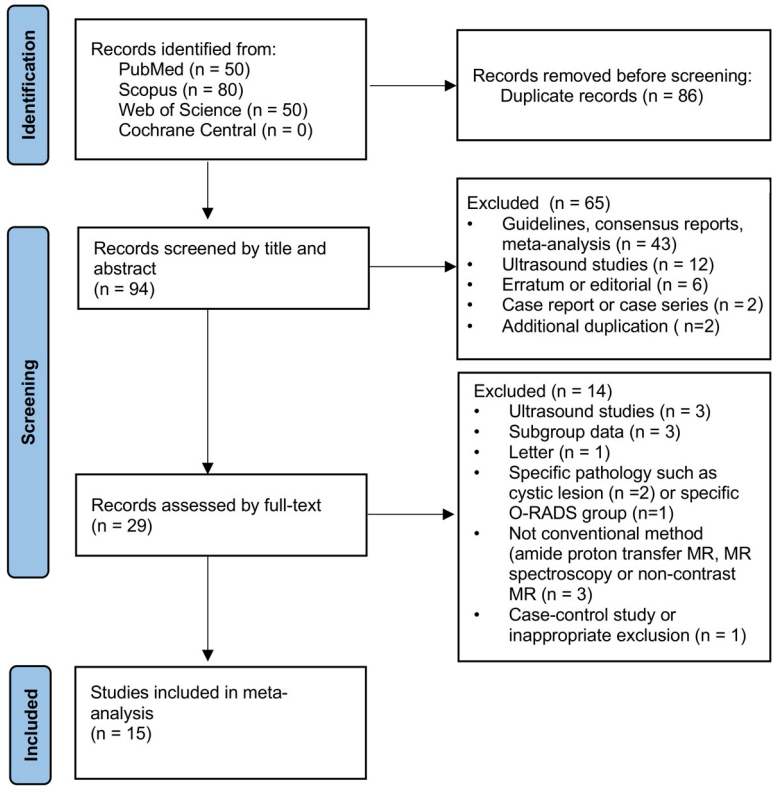


Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-analyses flowchart for selection of eligible studies. O-RADS, Ovarian-Adnexal Reporting and Data System; MR, magnetic resonance.

Table 1. The characteristics of the included studies

Author	Publication year	Screening period	Number of patients	Number of lesions	Percentage of the borderline or malignant lesions (%)	Mean or median age (years)	Reference standard
Aslan and Tosun[2]	2023	Jan 2018-June 2020	200	237	11.8	56.3	Pathology or 24-month follow-up
Bang et al.[3]	2022	Jan 2014-July 2020 and Jan 2010-July 2020	110	110	54.6	50.8	Pathology
Basu et al.[4]	2022	April 2020-June 2021	42	46	28.3	35.9	Pathology or 4-month follow-up
Campos et al.[5]	2023	Mar 5, 2013-Dec 31, 2021	227	269	11.2	40	Pathology or 24-month follow-up
Chen et al.[6]	2023	Jan 2017-Aug 2021	309	327	11.8	-	Pathology
Crestani et al.[7]	2020	2014-2018	26	26	65.4	57	Pathology
Elshetry et al.[8]	2023	April 2020-Sep 2021	90	116	38.8	39.4	Pathology or 12-month follow-up
Guo et al.[9]	2022	July 2017-June 2020	54	56	14.3	37	Pathology and median 1.2-year follow-up
Hottat et al.[10]	2022	Jan 2015-April 2020	163	201	28.9	51	Pathology
Manganaro et al.[11]	2023	Jan 2015-June 2022	140	172	52.9	48.7	Pathology or 12-month follow-up
Pereira et al.[12]	2022	Feb 2014-Dec 2020	226	287	31.4	-	Pathology or 12-month follow-up
Thomassin-Naggara et al.[16]	2020	Mar 1, 2013-Mar 31, 2016	1,130	1502	13.5	49	Pathology or 24-month follow-up
Wang et al.[13]	2023	May 2017-July 2022	240	278	25.2	42	Pathology or 6–12 months of follow-up
Woo et al.[14]	2023 (online ahead of print)	April 2021-Aug 2022	119	119	17.6	50	Pathology or ≥6-month follow-up
Wu et al.[15]	2023	Jan 2018-Mar 2022	308	362	11.6	42.1	Pathology

and 4.66 (P = 0.324), respectively]. However, substantial heterogeneity was observed for specificity [I^2 values were 64.7% for the overall group and 66.20% and 62.4% for the studies with low risk and high risk of selection bias, respectively. The corresponding Cochran's Q statistics (P values) were 39.66 (P < 0.001), 26.63 (P = 0.002), and 10.6 (P = 0.031), respectively].

Meta-regression analysis revealed that the pooled sensitivity was significantly different for the studies with low risk and high risk of bias; the sensitivity values were slightly, but significantly, lower for the studies with low risk of bias compared with those with a high risk of bias [the relative pooled sensitivity for low risk versus high risk of bias studies was 0.954 (95% CI 0.911–0.999), P = 0.043]. Therefore, the pooled sensitivity values are given separately for the studies with low and high risk of bias (Figure 3), and they were 93.0% (95% CI 89.1%–95.5%, with high certainty of evidence) for the studies with low risk of bias, and 97.5% (95% CI 91.3%–99.3%) for the studies with high risk of bias. The pooled specificities were not significantly different for the studies with low and high risk of bias [the relative specificity for the studies with low vs. high risk of bias was 1.014 [(95% CI 0.930–1.106); P = 0.752]. The pooled specificity for the overall study population was 90.4% (95% CI 86.6%–93.2%, with moderate certainty of evidence due to high unexplained heterogeneity; Figure 3). The model performance did not increase with the inclusion of the variables of mean age and proportion of borderline or malignant lesions into the regression model (P = 0.232).

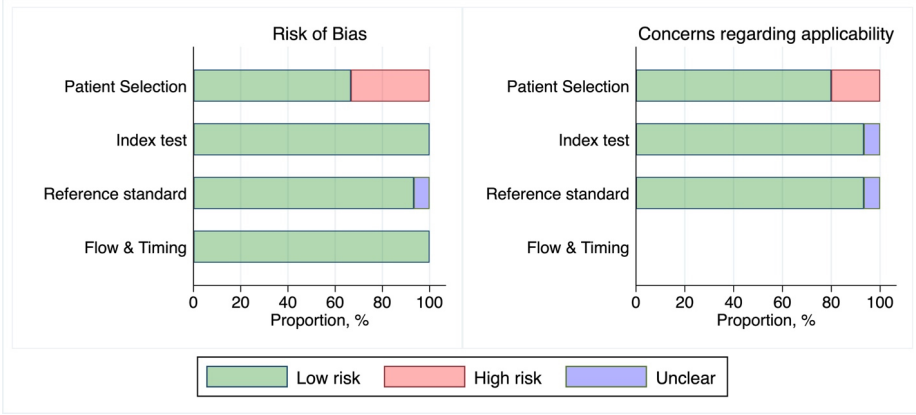


Figure 2. Methodological quality assessment according to the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

The SROC plot is presented in Figure 4 (the SROC plot with confidence and prediction intervals is given in Supplementary Figure 1). The plot shows that the diagnostic performance of the O-RADS system is very high (the point estimate is very close to the upper left corner of the SROC plot). Additionally, the plot reveals that the diagnostic performance is slightly lower for the studies with a low risk of bias compared with those with a high risk of bias [area under the curve 0.97 (95% CI 0.95–0.98); P < 0.001 vs. 0.99 (95% CI 0.97–0.99); P < 0.001, respectively], probably

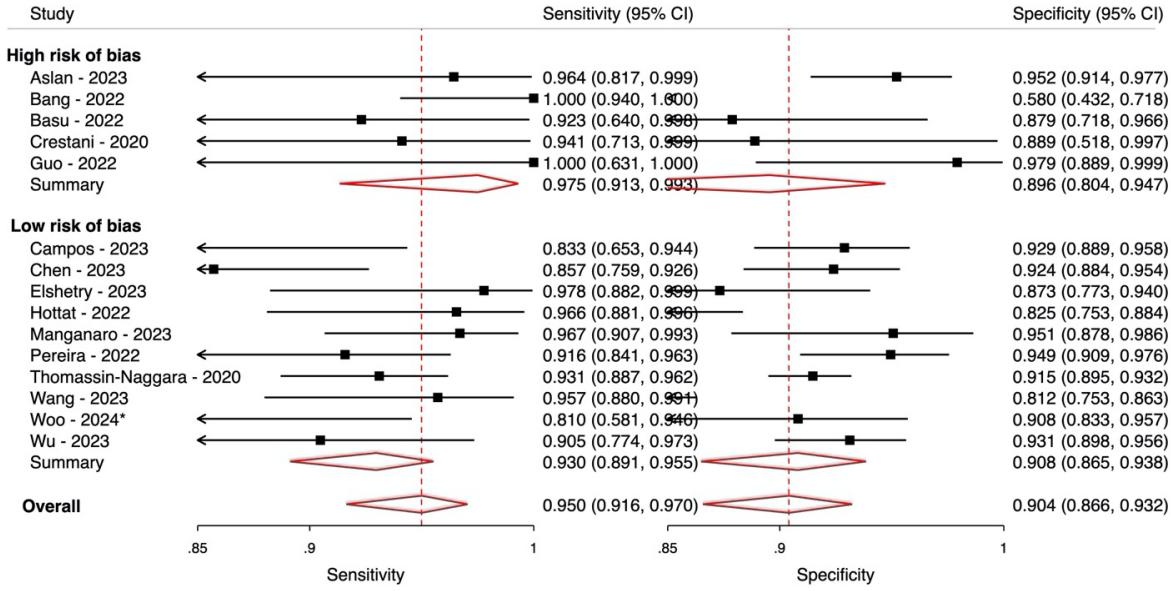


Figure 3. Forest plot of the pooled sensitivity and specificity. *Online publication in 2023, ahead of print.

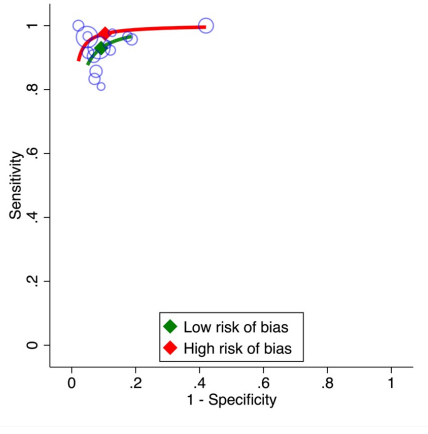


Figure 4. Summary receiver operating characteristics curve for the diagnostic performance of the Ovarian-Adnexal Reporting and Data System scoring. The blue circles represent individual studies, with their sizes proportional to the respective sample sizes. The red and green diamonds denote the point estimates (summary points), while the red and green solid lines illustrate the summary curves for studies with high and low risk of bias, respectively. For a more detailed depiction, including the confidence interval and prediction interval, please refer to Supplementary Figure 1.

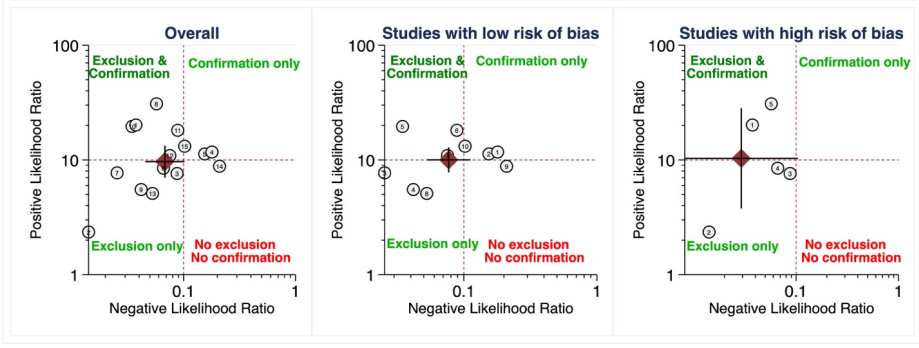


Figure 5. Likelihood matrix shows the pooled estimate (the diamond) and 95% confidence intervals of the negative and positive likelihood ratios, and exclusion and/or confirmation potential of the Ovarian-Adnexal Reporting and Data System scoring for borderline or malignant lesions.

due to lower pooled sensitivity in the low-risk bias group. However, Supplementary Figure 1 reveals that the precision is higher for the studies with a low risk of bias.

The pooled positive and negative LR values are provided in Supplementary Table 2. The LR matrix plot (Figure 5) indicates that the O-RADS system is more valuable for ruling out borderline or malignant lesions. In the overall population (Figure 5, left panel), the upper limit of the 95% CI of the negative LR is just at the cut-off limit of 0.1 [negative LR 0.07 (95% CI 0.05–0.10)]. A similar finding was observed for those with a high risk of bias but with a wider CI (Figure 5, right panel, Supplementary Table 2). Although the point estimate of the negative LR for the studies with low risk of bias was in the rule-out zone, the CI slightly crossed the cut-off value of 0.1 [negative LR for the low-risk group was 0.08 (95% CI 0.05–0.11)]. The point estimate of the pooled positive LR value was around the cut-off value of 10, with a lower limit of 95% CI >5, except for the value obtained from the studies with a high risk of bias. This suggests that the ruling-in potency of O-RADS scoring is moderate. The Fagan's nomogram demonstrates obtaining the post-test probability of having borderline or malignant lesion depending on the positive (O-RADS 4 or 5) or negative (O-RADS <4) test result (Figure 6).

Deeks' funnel plot indicates that there is no concern for publication bias (P = 0.812; Figure 7).

Discussion

This systematic review and meta-analysis show that 1) the pooled sensitivity of O-RADS MRI scores ≥4 in diagnosing borderline or malignant adnexal tumors is high and varies slightly according to whether the study population has a low or high risk of patient selection bias [the sensitivity is slightly, but significantly, lower in the low-risk of bias

group (93.0% vs. 97.5%)]; 2) the pooled specificity of the O-RADS MRI score is 90.4% in the overall population with no significant difference between the studies with low risk and high risk of selection bias, and 3) using the cut-off value of ≥4, the O-RADS MRI scores is valuable in ruling out the borderline or malignant lesions, although the ruling-in potency is relatively lower.

Ovarian cancers are estimated to be responsible for 5% of cancer-related deaths in women, with a 5-year survival rate of 50%.[25]

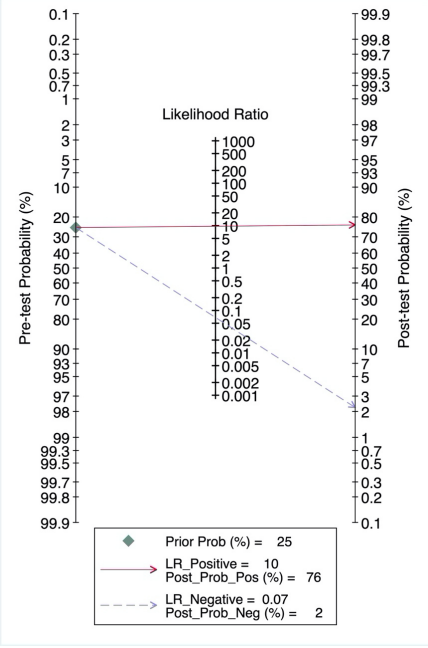


Figure 6. Fagan's nomogram for the Ovarian-Adnexal Reporting and Data System (O-RADS) scoring. The green diamond on the pre-test probability line (on the left side) represents the overall pre-test probability (25%) obtained from this meta-analysis. Utilizing the pooled likelihood ratio values, the solid red arrow and the dashed grey arrow indicate the post-test probability of having a borderline or malignant lesion when the test is positive (O-RADS 4 or 5) or negative (O-RADS <4), respectively. LR, likelihood ratio.

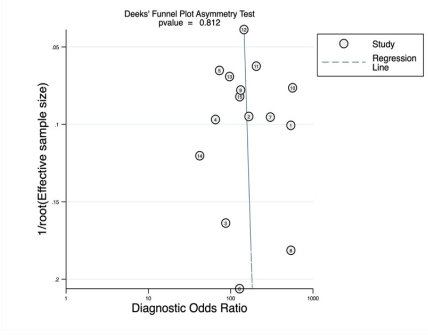


Figure 7. Deeks' plot for publication bias.

Ultrasonography is the first-line diagnostic method due to its low cost and wide availability. MRI is a better diagnostic method in terms of characterization and determining the origin of adnexal masses. To have a similar lexicon between radiologists and clinicians and for accurate referral of patients for surgical treatment, the ACR developed a system- O-RADS- for the characterization of adnexal masses.[1] After introducing the O-RADS system, several studies assessing its diagnostic performance have been published. In this study, a systematic review and meta-analysis of these studies were conducted to obtain updated information along with consideration of the risk of selection bias for each study.

The present study demonstrates that the heterogeneity between the studies for diagnostic sensitivity is not high, which implies that the results of the studies among the borderline or malignant lesions are consistent. On the other hand, the heterogeneity between the studies for diagnostic specificity is high. This implies that the consistency of the O-RADS system for benign lesions is relatively low, particularly for studies with a high risk of patient selection bias. High heterogeneity for specificity was also demonstrated in a previous meta-analysis,[18] but that included a lower number of studies and did not consider the risk of bias while pooling the results.

Both the sensitivity and specificity of the O-RADS system in discriminating benign lesions and borderline or malignant lesions are high, although the sensitivity is higher than the specificity. The pooled sensitivity varies for those with or without a high risk of patient selection bias, and studies with a low risk of bias have a lower, but acceptable, pooled sensitivity (93.0% vs. 97.5%). Specificity was also high, but similar for the studies with low or high risk of selection bias. Consistent with these findings, the SROC plot shows that the O-RADS system has high diagnostic performance (discrimination) for borderline or malignant lesions. This suggests that the O-RADS system is a good tool for referring patients to surgery. The SROC plot (Supplementary Figure 1) also shows that the precision (based on the 95% CI and prediction interval) is very high for the studies with a low risk of bias but is relatively lower for the studies with a high risk of bias.

Because of the high sensitivity, O-RADS MRI scoring is valuable for ruling out borderline or malignant lesions. This is supported by the LR matrix plot. It is generally accepted that a negative LR value of <0.1 indicates that the test is valuable in ruling out the disease, and a positive LR value of >10 indicates the test is valuable in ruling in the disease,[26,27] although they are arbitrarily chosen cut-off values. Furthermore, negative LR values of 0.1–0.2 and positive LR values of 5–10 indicate that the test is moderately effective in ruling out and ruling in the disease, respectively. In the present study, the upper limit of the 95% CI of the negative LR value was just at the cut-off value of 0.1 in the overall population and in the analysis of the studies with a high risk of bias, which suggests O-RADS MRI is good at in excluding the disease. For the studies with a low risk of bias, although the upper limit of 95% CI for the negative LR slightly crossed the cut-off value (negative LR value 0.08, 95% CI 0.05–0.11), the ruling-out potential was largely preserved. The point estimates of the positive LR values were around the cut-off value of 10, and although the CI crossed the cut-off value of 10, the lower limit was >5 for the overall population and those with a low risk of selection bias (Supplementary Table 2). This finding suggests the O-RADS MRI score is moderately effective in ruling in the disease.

In the EURAD study, the prospective European multicenter cohort, misclassified cases were assessed in terms of three types of error: errors caused by technical limitations, inadequate experience (perceptual error), or interpretive errors.[28] The interpretive error was found to be the most common cause of the misclassification, which was mostly due to rating benign lesions as O-RADS 4 or 5 (false positive result). Even if some of the false positive results were caused by a concern for missing the malignancy, they demonstrated that the misclassification was substantially reduced with strict application of O-RADS scoring. The false positive result is associated with low specificity. In the present study, the heterogeneity between the studies was high for the specificity. Additionally, compared with sensitivity, specificity was relatively low. This may suggest a problem with the rating of benign lesions. Therefore, approaches to increase specificity and reduce potential heterogeneity in the interpretation of the benign lesions may reduce unnecessary surgical procedures by keeping the false positive rate low. This may be obtained by reducing interpretive errors by applying the O-RADS scoring meticulously and by increasing the awareness of some lesions that may be misclassified. Thomassin-Naggara et al.[28] discussed these lesions in their article and underlined the importance of the difference between a solid lesion and a solid compo-

nent for correct classification. Another factor for correct classification is the availability of technically adequate MRI images, which has been discussed elsewhere.[29] In addition, some refinement in the O-RADS system may improve its diagnostic value. Several methods seem promising in increasing the diagnostic performance of the O-RADS system. Wengert et al.[30] showed that time-intensity curve analysis was superior to visual assessment and improved the specificity. Furthermore, diffusion-weighted imaging improves the diagnostic performance of the O-RADS MRI system.[10] Application of these methods may reduce false positive results by increasing specificity and may also increase its ruling-out potential further by increasing the sensitivity.

The present study has several limitations. First, as in many meta-analyses, data were extracted from published manuscripts; therefore, individual participant data were not available. Although it is very difficult to obtain, individual participant data analysis provides more reliable information and may provide detailed reasons for heterogeneity. Second, we did not analyze the data based on the readers' experience; other confounding factors may also affect the results. However, the relatively low number of studies precludes taking many factors into consideration, especially if individual data are not available. Third, we aimed to assess the "intrinsic" diagnostic performance of O-RADS MRI scoring; therefore, cancer antigen 125 levels, or other factors such as menopausal status or family history that may be used to assess the pre-test probability of the malignancy, were not used in the analysis. Instead, we preferred to provide Fagan's nomogram, in which the pre-test probability obtained by any marker or clinical predictors can be combined with the "intrinsic" performance of the O-RADS score to obtain the post-test probability. The present analysis also has some advantages, such as including new studies, and, in contrast to recent meta-analyses, assessing the diagnostic performance and providing visual information about the ruling-in and ruling-out potential according to the risk of bias.

In conclusion, O-RADS MRI scoring is valuable in ruling out borderline or malignant lesions, while the ruling-in potency is moderate. Patient selection bias affects diagnostic sensitivity, leading to a higher sensitivity compared with the sensitivity obtained from the studies with a low risk of bias. The high between-study heterogeneity observed for specificity suggests the need for improvement in the consistent characterization of the benign lesions to reduce false positive rates.

Acknowledgement

Thanks to Dr. Cansu Öztürk for data handling and critical appraisal of the literature, and to Dr. Mustafa Kılıçkap for conducting statistical analyses and constructive comments.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Reinhold C, Rockall A, Sadowski EA, et al. Ovarian-Adnexal Reporting Lexicon for MRI: A White Paper of the ACR Ovarian-Adnexal Reporting and Data Systems MRI Committee. *J Am Coll Radiol.* 2021;18(5):713-729. [CrossRef]
2. Aslan S, Tosun SA. Diagnostic accuracy and validity of the O-RADS MRI score based on a simplified MRI protocol: a single tertiary center retrospective study. *Acta Radiol.* 2023;64(1):377-386. [CrossRef]
3. Bang JI, Kim JY, Choi MC, Lee HY, Jang SJ. Application of multimodal imaging biomarker in the differential diagnosis of ovarian mass: integration of conventional and molecular imaging. *Clin Nucl Med.* 2022;47(2):117-122. [CrossRef]
4. Basu A, Pame M, Bhuyan R, Roy DK, James VM. Diagnostic Performance of O-RADS MRI scoring system for the assessment of adnexal masses in routine clinical radiology practice-a single tertiary centre prospective cohort study. *J Clin Diagn Res.* 2022;16(4):TC11-TC16. [CrossRef]
5. Campos A, Villermain-Lécolier C, Sadowski EA, et al. O-RADS scoring system for adnexal lesions: diagnostic performance on TVUS performed by an expert sonographer and MRI. *Eur J Radiol.* 2023;169:111172. [CrossRef]
6. Chen T, Qian X, Wei C, et al. The consistency and application value of MRI-based ovarian-adnexal reporting and data system in the diagnosis of ovarian adnexal masses. *Chinese Journal of Radiology.* 2023;57(12):282-287. [CrossRef]
7. Crestani A, Theodore C, Levaillant JM, et al. Magnetic resonance and ultrasound fusion imaging to characterise ovarian masses: a feasibility study. *Anticancer Res.* 2020;40(7):4115-4121. [CrossRef]
8. Elshetry ASF, Hamed EM, Frere RAF, Zaid NA. Impact of adding mean apparent diffusion coefficient (ADCmean) measurements to O-RADS MRI scoring for adnexal lesions characterization: a combined O-RADS MRI/ADCmean approach. *Acad Radiol.* 2023;30(2):300-311. [CrossRef]
9. Guo Y, Phillips CH, Suarez-Weiss K, et al. Interreader agreement and intermodality concordance of O-RADS US and MRI for assessing large, complex ovarian-adnexal cysts. *Radiol Imaging Cancer.* 2022;4(5):e220064. [CrossRef]
10. Hottat NA, Badr DA, Van Pachterbeke C, et al. Added value of quantitative analysis of diffusion-weighted imaging in Ovarian-Adnexal Reporting and Data System Magnetic Resonance Imaging. *J Magn Reson Imaging.* 2022;56(1):158-170. [CrossRef]
11. Manganaro L, Ciulla S, Celli V, et al. Impact of DWI and ADC values in Ovarian-Adnexal Reporting and Data System (O-RADS) MRI score. *Radiol Med.* 2023;128(5):565-577. [CrossRef]
12. Pereira PN, Yoshida A, Sarian LO, Barros RHO, Jales RM, Derchain S. Assessment of the performance of the O-RADS MRI score for the evaluation of adnexal masses, with technical notes. *Radiol Bras.* 2022;55(3):137-144. [CrossRef]
13. Wang T, Cui W, Nie F, et al. Comparative study of the efficacy of the Ovarian-Adnexa Reporting and Data System Ultrasound Combined With Contrast-Enhanced Ultrasound and the ADNEX MR scoring system in the diagnosis of adnexal masses. *Ultrasound Med Biol.* 2023;49(9):2072-2080. [CrossRef]
14. Woo S, Andrieu PC, Abu-Rustum NR, et al. Bridging communication gaps between radiologists, referring physicians, and patients through standardized structured cancer imaging reporting: the experience with female pelvic MRI assessment using O-RADS and a simulated cohort patient group. *Acad Radiol.* 2024;31(4):1388-1397. [CrossRef]
15. Wu M, Tang Q, Cai S, et al. Accuracy and reproducibility of the O-RADS MRI risk stratification system based on enhanced non-DCE MRI in the assessment of adnexal masses. *Eur J Radiol.* 2023;159:110670. [CrossRef]
16. Thomassin-Naggara I, Poncelet E, Jalaguier-Coudray A, et al. Ovarian-Adnexal Reporting Data System Magnetic Resonance Imaging (O-RADS MRI) score for risk stratification of sonographically indeterminate adnexal masses. *JAMA Netw Open.* 2020;3(1):e1919896. [CrossRef]
17. Rizzo S, Cozzi A, Dolciami M, et al. O-RADS MRI: A systematic review and meta-analysis of diagnostic performance and category-wise malignancy rates. *Radiology.* 2023;307(1):e220795. [CrossRef]
18. Zhang Q, Dai X, Li W. Systematic review and meta-analysis of O-RADS Ultrasound and O-RADS MRI for risk assessment of ovarian and adnexal lesions. *AJR Am J Roentgenol.* 2023;221(1):21-33. [CrossRef]

19. Shamseer L, Moher D, Clarke M, et al. Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*. 2015;350:g7647. Erratum in: *BMJ*. 2016;354:i4086. [CrossRef]
20. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann Intern Med*. 2011;155(8):529-536. [CrossRef]
21. Schünemann H BJ, Guyatt G, Oxman A. GRADE handbook. Accessed May, 24 2024. [CrossRef]
22. Yang WT, Parikh JR, Stavros AT, Otto P, Maislin G. Exploring the negative likelihood ratio and how it can be used to minimize false-positives in breast imaging. *AJR Am J Roentgenol*. 2018;210(2):301-306. [CrossRef]
23. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: the case of meta-analyses of diagnostic accuracy. *Stat Med*. 2014;33(16):2701-2717. [CrossRef]
24. Deeks JJ, Macaskill P, Irwig L. The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed. *J Clin Epidemiol*. 2005;58(9):882-893. [CrossRef]
25. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin*. 2023;73(1):17-48. [CrossRef]
26. Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. *Phys Ther*. 2001;81(9):1546-1564. [CrossRef]
27. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet*. 2005;365(9469):1500-1505. [CrossRef]
28. Thomassin-Naggara I, Belghitti M, Milon A, et al. O-RADS MRI score: analysis of misclassified cases in a prospective multicentric European cohort. *Eur Radiol*. 2021;31(12):9588-9599. [CrossRef]
29. Nougaret S, Lakhman Y, Bahadir S, Sadowski E, Thomassin-Naggara I, Reinhold C. Ovarian-Adnexal Reporting and Data System for Magnetic Resonance Imaging (O-RADS MRI): Genesis and Future Directions. *Can Assoc Radiol J*. 2023;74(2):370-381. [CrossRef]
30. Wengert GJ, Dabi Y, Kermarrec E, et al. O-RADS MRI classification of indeterminate adnexal lesions: time-intensity curve analysis is better than visual assessment. *Radiology*. 2022;303(3):566-575. Erratum in: *Radiology*. 2022;303(2):E28. [CrossRef]

Supplementary Table 1. Assessment of the methodological quality of each study according to the QUADAS-2 tool

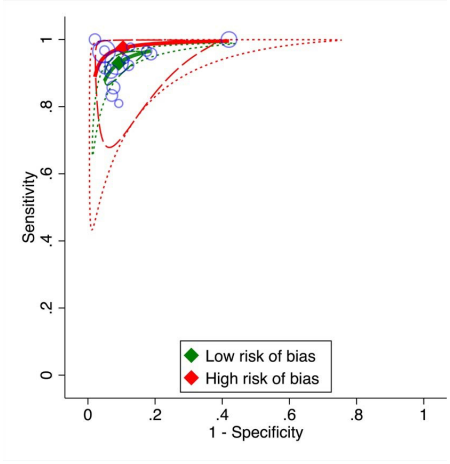
	Risk of bias				Concerns regarding applicability		
Study	Patient selection	Index test	Reference test	Flow & timing	Patient selection	Index test	Reference test
Aslan and Tosun[2] - 2023	High*	Low	Low	Low	Low	Unclear	Low
Bang et al.[3] - 2022	High**	Low	Low	Low	High	Low	Low
Basu et al.[4] - 2022	High§	Low	Unclear	Low	High	Low	Unclear
Campos et al.[5] - 2023	Low	Low	Low	Low	Low	Low	Low
Chen et al.[6] - 2023	Low	Low	Low	Low	Low	Low	Low
Crestani et al.[7] - 2020	High§§	Low	Low	Low	High	Low	Low
Elshetry et al.[8] - 2023	Low	Low	Low	Low	Low	Low	Low
Guo et al.[9]- 2022	High‡	Low	Low	Low	Low	Low	Low
Hottat et al.[10] - 2022	Low	Low	Low	Low	Low	Low	Low
Manganaro et al.[11] - 2023	Low	Low	Low	Low	Low	Low	Low
Pereira et al.[12] - 2022	Low	Low	Low	Low	Low	Low	Low
Thomassin-Naggara et al.[16] - 2020	Low	Low	Low	Low	Low	Low	Low
Wang et al.[13] - 2023	Low	Low	Low	Low	Low	Low	Low
Woo et al.[14]- 2023 (ahead of print)	Low	Low	Low	Low	Low	Low	Low
Wu et al.[15] - 2023	Low	Low	Low	Low	Low	Low	Low

*, Simplified method and exclusion of <3 cm cysts; **, includes patients underwent PET/CT; §, Non-probability sampling and 4-month of follow-up; §§, includes a sub-population who underwent surgery; ‡, included patients with >5 cm cystic lesions; QUADAS-2, Quality Assessment of Diagnostic Accuracy Studies-2.

Supplementary Table 2. The negative and positive likelihood ratios for the overall population and for the studies with low or high-risk of bias

	Negative likelihood ratio and 95% CI	Positive likelihood ratio and 95% CI
Studies with low risk of bias	0.08 (0.05 – 0.11)	10.0 (7.7 – 12.9)
Studies with high risk of bias	0.03 (0.01 – 0.10)	10.3 (3.8 – 28.3)
Overall population	0.07 (0.05 – 0.10)	9.7 (7.0 – 13.3)

CI, confidence interval



Supplementary Figure 1. Summary receiver operating characteristics curve for the diagnostic performance of the O-RADS scoring. The blue circles represent each study, with their sizes proportional to the sample size of the respective study. The red and green diamonds depict the point estimates (summary points), while the red and green solid lines illustrate the summary curve for the studies with high and low-risk of bias, respectively. Correspondingly, the red and green dashed lines represent the confidence interval, and the red and green dotted lines indicate the prediction interval for the pooled estimates of studies with high and low-risk of bias, respectively. O-RADS, Ovarian-Adnexal Reporting and Data System.