GYNECOLOGY

External validation of ultrasound-based models for differentiating between benign and malignant adnexal masses: a nationwide prospective multicenter study (IOTA phase 6)

Francesca Moro, PhD; Marina Momi, MD; Ashleigh Ledger, MD; Lasai Barreñada, MD; Jolien Ceusters, PhD; Davide Sturla, MD; Elisa Mor, MD; Letizia Fornari, MD; Floriana Mascilini, MD; Francesca Ciccarone, MD; Federica Pozzati, MD; Wouter Froyman, PhD; Ben Van Calster, PhD; Tom Bourne, PhD; Dirk Timmerman, PhD; Anna Fagotti, PhD; Lil Valentin, PhD; Antonia Carla Testa, PhD; IOTA 6 Collaborators*

BACKGROUND: The diagnostic performance of the IOTA methods, the O-RADS lexicon, and the RMI has been validated in prospective and retrospective studies, but most validation studies tested the performance in the hands of experienced ultrasound examiners.

OBJECTIVE: To prospectively validate the performance of the Risk of Malignancy Index, the International Ovarian Tumor Analysis Simple Rules Risk Model, the International Ovarian Tumor Analysis Assessment of Different NEoplasias in the adneXa model, and the International Ovarian Tumor Analysis 2-step strategy across different types of ultrasound centers in Italy. A retrospective post hoc analysis estimates malignancy prevalence in Ovarian-Adnexal Reporting and Data System risk groups when using the 2-step strategy or the Ovarian-Adnexal Reporting and Data System lexicon.

STUDY DESIGN: This is a multicenter prospective observational study including regional referral centers and district hospitals in Italy.

METHODS: Consecutive patients with an adnexal mass examined with ultrasound by an International Ovarian Tumor Analysis—certified gynecologist with different levels of expertise were included, provided they underwent surgery $<$180 days after the inclusion scan. Ultrasound examination was performed transvaginally or transrectally and was supplemented with an abdominal scan when necessary. Reference standard was the histology of the adnexal mass following surgical removal. Discrimination (area under the receiver operating characteristic curve), calibration, and clinical utility were assessed to illustrate the diagnostic performance of the methods. For the retrospective post hoc analysis, we report the prevalence of malignancy in the Ovarian-Adnexal Reporting and Data System risk groups (Ovarian-Adnexal Reporting and Data System 2: risk of malignancy $<$1%; Ovarian-Adnexal Reporting and Data System 3: risk of malignancy 1% to $<$10%; Ovarian-Adnexal Reporting and Data System 4: risk of malignancy 10% to $<$50%; Ovarian-Adnexal Reporting and Data System 5: risk of malignancy $\geq$50%), with the Ovarian-Adnexal Reporting and Data System risk group assigned using either the 2-step strategy or the Ovarian-Adnexal Reporting and Data System lexicon.

RESULTS: Between May 2017 and March 2020, 1431 patients were enrolled from 21 Italian centers (10 oncological and 11 nononcological). Based on histology, 995 (69.5%) tumors were benign and 436 (30.5%) were malignant (115, 8.0% borderline; 263, 18.4% primary invasive; 58, 4.1% metastatic tumors).

For Risk of Malignancy Index, the area under the receiver operating characteristic curve was 0.85 (95% confidence interval, 0.81 to 0.87), whereas for all International Ovarian Tumor Analysis models (Simple Rules Risk Model, Assessment of Different NEoplasias in the adneXa with and without CA125, and 2-step strategy with and without CA125), the area under the receiver operating characteristic curves ranged from 0.91 (95% confidence interval, 0.88—0.93) to 0.92 (95% confidence interval, 0.89—0.94). All International Ovarian Tumor Analysis models demonstrated a higher net benefit than Risk of Malignancy Index across risk thresholds (exchange rates) from 1% to 50%. All International Ovarian Tumor Analysis models slightly underestimated the risk of malignancy, but Simple Rules Risk Model showed the least degree of underestimation.

The prevalence of malignancy and the corresponding 95% confidence interval in the 4 Ovarian-Adnexal Reporting and Data System risk categories, as calculated using the 2-step strategy and Ovarian-Adnexal Reporting and Data System lexicon, were 0.97% (95% confidence interval, 0.4—2.6) and 1.2% (95% confidence interval, 0.5—2.9) for Ovarian-Adnexal Reporting and Data System 2, 7.2% (95% confidence interval, 5.0—10.3) and 6.0% (95% confidence interval, 3.6—9.6) for Ovarian-Adnexal Reporting and Data System 3, 37.9% (95% confidence interval, 32.4—43.8) and 27.8% (95% confidence interval, 23.6—32.5) for Ovarian-Adnexal Reporting and Data System 4, and 84% (95% confidence interval, 79.8—87.4) and 83.1% (95% confidence interval, 79.0—86.6) for Ovarian-Adnexal Reporting and Data System 5.

CONCLUSION: Risk of Malignancy Index had lower ability than the International Ovarian Tumor Analysis models to distinguish between benign and malignant adnexal tumors in patients examined by either expert or nonexpert ultrasound operators in Italy. All the International Ovarian Tumor Analysis models—including Simple Rules Risk Model, Assessment of Different NEoplasias in the adneXa, and the 2-step strategy with or without CA125—had similar ability. The prevalence of malignancy in each of the 4 Ovarian-Adnexal Reporting and Data System risk categories closely matched the assigned malignancy risk regardless of whether the 2-step strategy or Ovarian-Adnexal Reporting and Data System lexicon was used.

Key words: ovarian neoplasm, ultrasonography, validation study

Cite this article as: Moro F, Momi M, Ledger A, et al. External validation of ultrasound-based models for differentiating between benign and malignant adnexal masses: a nationwide prospective multicenter study (IOTA phase 6). Am J Obstet Gynecol 2025;XXX:XX—XX.

0002-9378

https://doi.org/10.1016/j.ajog.2025.07.017

Click Supplemental Materials under article title in Contents at ajog.org

AJOG at a Glance

Why was this study conducted?

Our study is a prospective national multicenter study that validates the International Ovarian Tumor Analysis models in the hands of ultrasound examiners with different levels of ultrasound expertise.

Key findings

The International Ovarian Tumor Analysis models have the ability to discern between benign and malignant adnexal masses regardless of whether the operator is an expert or not.

What does this add to what is known?

The prevalence of malignancy in each of the 4 Ovarian-Adnexal Reporting and Data System risk groups closely matched the assigned malignancy risk regardless of whether the 2-step strategy or Ovarian-Adnexal Reporting and Data System lexicon was used.

Introduction

Ovarian cancer is the leading cause of death in women diagnosed with gynecological cancers.[1] Ovarian cancers should be treated by gynecological oncological surgeons to optimize outcome.[2–4] Correct preoperative characterization of adnexal masses is essential to decide on optimal management: clinical and ultrasound follow-up, surgery in a local center, or referral to an oncology center.[5,6] Transvaginal ultrasound is the first-line method for characterizing adnexal masses. If performed by an expert, subjective assessment of the ultrasound images is the best method for distinguishing benign from malignant masses.[7–9] For less experienced ultrasound examiners, there are other methods. The Risk of Malignancy Index (RMI) is a scoring system using clinical and ultrasound information that can be used to estimate the likelihood of an ovarian mass being malignant.[10] In some European countries, RMI is widely used to triage women with an adnexal mass for referral to an oncological center.[11–15]

The International Ovarian Tumor Analysis (IOTA) group was founded in 1999 with the aim of establishing a standardized terminology and methodology for the assessment of adnexal masses with ultrasound. The group has developed several ultrasound-based methods to discriminate between benign and malignant adnexal masses. These include the Benign Descriptors (BDs), which consist of 4 simple ultrasound criteria; if any of them apply, the mass is classified as benign.[16,17] Another method is the Simple Rules, which include 5 benign and 5 malignant ultrasound features. If at least one benign feature is present and no malignant features are observed, the mass is classified as benign; conversely, if one or more malignant features are present and no benign features apply, the mass is classified as malignant. If both benign and malignant features apply, or if none of the 10 features is present, the mass cannot be classified using the Simple Rules (inconclusive result).[18] In addition, 4 mathematical models have been developed: Logistic Regression Model 1, Logistic Regression Model 2, the Simple Rules Risk Model (SRRisk), and the Assessment of Different NEoplasias in the adneXa (ADNEX).[19–21] Among these, ADNEX is the most extensively validated model.[22] ADNEX is a multinomial logistic regression model that incorporates 3 clinical and 6 ultrasound variables to calculate the probability of 5 types of tumor: benign, borderline, stage I primary invasive ovarian malignancy, stage II to IV primary invasive ovarian malignancy, or a secondary metastatic tumor.[21] Currently, the IOTA group recommends the 2-step strategy, which means first applying the BDs; if none of them apply, the ADNEX model is used.[17]

In 2020, the IOTA group and the American College of Radiology published together a consensus guideline on the ultrasound Ovarian-Adnexal Reporting and Data System (O-RADS). O-RADS categorizes adnexal masses into 6 risk groups ranging from a normal ovary to high risk of malignancy. According to the consensus statement, malignancy risk can be estimated using either the ADNEX model or the ultrasound examiner's interpretation of ultrasound findings based on the O-RADS lexicon.[23]

The diagnostic performance of the IOTA methods, the O-RADS lexicon, and the RMI has been validated in prospective and retrospective studies, but most validation studies tested the performance in the hands of experienced ultrasound examiners.[17,24–35] No prospective study included nonexpert ultrasound examiners and few included examiners with different levels of expertise.[36–38]

The aims of the study are (1) to prospectively validate the diagnostic performance of RMI, SRRisk, ADNEX, and the IOTA 2-step strategy in different types of ultrasound centers in Italy both overall and in relevant subgroups and to assess the ability of the BDs to correctly classify adnexal masses as benign; (2) to validate the performance of the Simple Rules and subjective assessment overall and in relevant subgroups; and (3) to conduct a retrospective post hoc analysis estimating the prevalence of malignancy within the O-RADS risk groups, using the O-RADS lexicon and the 2-step strategy.

Methods

Study design and participants

This is an Italian multicenter prospective external validation study of ultrasound-based models to discriminate between benign and malignant adnexal masses. The protocol was approved by the Ethical Committee of the Fondazione Policlinico A. Gemelli, IRCCS (PROT 27665/16) and of each participating center (Appendix 1). Written informed consent was obtained from all patients. We report the study following the Transparent Reporting of a multivariable prediction model for

Individual Prognosis Or Diagnosis (TRIPOD) cluster guidelines.[39]

Consecutive patients with a known or suspected adnexal mass examined with ultrasound by an IOTA-certified gynecologist[40] (see below) and confirmed to have an adnexal mass judged not to be physiological were eligible for inclusion provided they were expected to undergo surgical removal of the mass. The patients were recruited between May 2017 and March 2020. Exclusion criteria were: patient's age <18 years, pregnant patients, patients with previous bilateral adnexectomy, patients examined in centers that recruited <10 patients, only transabdominal ultrasound performed, surgery performed more than 180 days after the ultrasound examination, and denial or withdrawal of informed consent.

Information on age, parity, menopausal status, and indication for the ultrasound examination was prospectively collected and information on type of hospital (private practice, local public hospital, regional public hospital, or university hospital), type of center (oncological vs nononcological), and type of ultrasound center (general gynecologic outpatient clinic or specialized ultrasound center). An oncological center was defined as a tertiary referral center with a dedicated gynecological oncology unit. Information on the ultrasound system used, ultrasound examiner's name, and level of ultrasound experience was also recorded. All ultrasound examinations were performed by IOTA-certified gynecologists. IOTA certification is obtained after participation in an IOTA certification course and passing an IOTA certification test (www.iotagroup.org).

The level of ultrasound experience was based on the number of gynecological scans in nonpregnant women that the examiner had performed at the start of the study. Low experience was defined (partly arbitrarily, partly based of the European Federation of Societies for Ultrasound in Medicine and Biology [EFSUMB] guidelines)[41] as <500 scans, intermediate experience as 500−5000 scans, and high experience as >5000 scans. We also recorded the level of expertise according to EFSUMB (level 1, 2, or 3).

A standardized transvaginal (or transrectal if vaginal was not possible) ultrasound examination including color or power Doppler examination was performed, supplemented with transabdominal ultrasound when necessary. The IOTA examination and measurement technique were used, and the ultrasound findings were described using the IOTA terminology.[42] Information on all the variables required for RMI, SRRisk, ADNEX, BDs, and Simple Rules were prospectively collected and recorded. Results of subjective assessment were recorded as benign, borderline, or malignant. The degree of diagnostic confidence (certainly benign, probably benign, uncertain, probably malignant or probably borderline, certainly malignant or certainly borderline) and the specific diagnosis suggested by the examiner and chosen from a list of predefined diagnoses were also recorded.

If more than one adnexal mass was present, only the one with the most complex ultrasound morphology was included in our statistical analysis. If the ultrasound morphology was similar in all masses, the largest one or the one most easily accessible with ultrasound was used in our statistical calculations. The management was decided by the referring clinician, who took into account clinical symptoms, ultrasound results based on subjective evaluation of the ultrasound images (ie, those reported in the clinical ultrasound report), and results of other imaging modalities (eg, computer tomography or magnetic resonance imaging), tumor markers, and patient's preference.

Reference standard was the histology of the adnexal mass following surgical removal within 180 days after the ultrasound examination. The histology of the surgically removed tumor was determined at the local center. Pathologists were blinded to ultrasound predictor variables and model predictions but might have received information on the subjective assessment by the ultrasound examiner when clinically relevant. The stage of malignant tumors was recorded using the classification of the International Federation of Gynecology and Obstetrics.[43]

Data collection was done through the web-based clinical data miner software.[44] A team of statisticians and ultrasound examiners performed data cleaning. A description of the models is provided in Appendix 2.

Statistical analysis and sample size

The statistical analyses were performed with R version 4.1.2. The adequacy of the sample size is discussed in Appendix 3. Despite it being strongly recommended to collect blood samples for the measurement of serum CA125 in all patients, CA125 results were missing in some patients. Missing CA125 values were imputed. We performed multiple imputations using the method of fully conditional specification. The multiple imputation procedure is described in the Appendix 4. To calculate the predictions for each model, we used the formula presented in the original publications. These calculations were performed after completion of the study and so were not available to guide the management of the patients. We report the percentage of tumors to which a BD applied and the outcome of masses to which a BD applied (pooled analysis).

We calculated center-specific area under the receiver operating characteristic curve (AUROC) to estimate the ability to discriminate between benign and malignant adnexal masses for RMI and the risk models (SRRisk, ADNEX, and IOTA 2-step strategy) and used meta-analysis to obtain the overall AUROC per model (Appendix 5). The heterogeneity between centers was assessed by calculating 95% prediction intervals. We also assessed calibration of the risk models by calculating observed over expected ratio (O:E). O:E is the ratio of the observed risk of having the outcome divided by the risk estimated by the model (O:E >1: model underestimates risk of malignancy; O:E <1: model overestimates risk of malignancy).[45,46] In addition, we constructed flexible calibration curves using locally estimated scatterplot smoothing.[47]

Clinical utility to decide which patients to refer for specialized oncological care was estimated by calculating the net benefit (NB) for risk thresholds between 1% and 50%.[48] We plotted the NB of the model at each risk threshold (exchange rate) in a decision curve. We also plotted a line for the treat all and treat none strategy (in this case to refer all or to refer none to an oncology center). A model is clinically useful if it is superior to both treat all and treat none.

We calculated sensitivity, specificity, positive and negative predictive values for subjective assessment and the Simple Rules (inconclusive cases classified as malignant), and for the risk models at risk of malignancy cutoffs between 1% and 50%. For RMI, we report classification performance at cutoffs 25, 100, 200, and 250. We calculated center-specific sensitivity and specificity and combined them using meta-analysis.[49]

To estimate the multinomial performance of ADNEX and the 2-step strategy, we computed the Polytomous Discrimination Index and calculated the AUROC for each pair of outcome categories using the conditional risk method.[50,51]

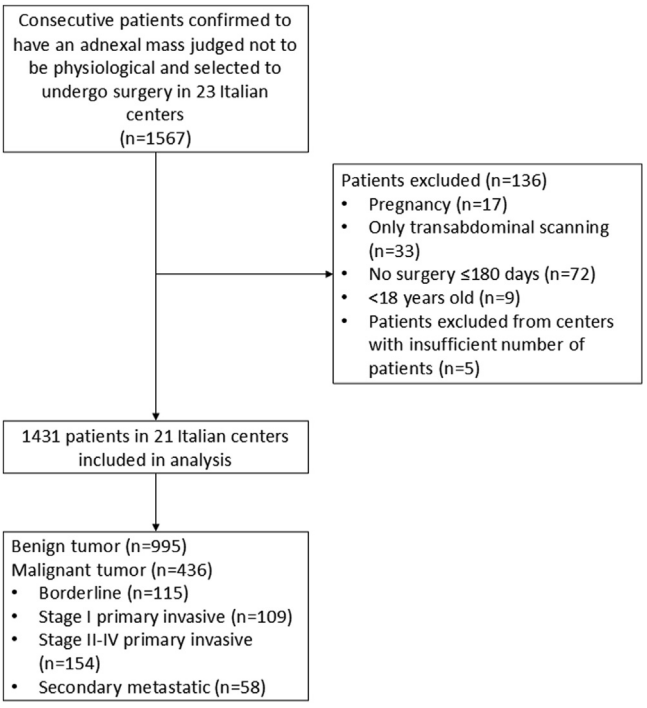
In a post hoc retrospective analysis, we calculated the prevalence of malignancy along with its 95% confidence interval (CI) in each of the O-RADS risk groups. Details about O-RADS calculations can be found in Appendix 2. To derive the risk groups using the O-RADS lexicon from the prospectively collected IOTA variables, we used the method described by Timmerman et al.[35]

More information about the statistical analyses and the meta-analyses are found in Appendix 5.

Subgroup analyses

We calculated AUROC and O:E ratio for prespecified subgroups based on menopausal status, type of center (oncology vs nononcology), and ultrasound examiner's experience (<500 scans performed, 500−5000 scans performed, and >5000 scans performed; EFSUMB Level 1, Level 2, and Level 3) using pooled data due to the small numbers in most centers.

FIGURE 1
Flowchart of the patients included for the analysis



Results

A total of 1567 patients were recruited from 23 Italian centers. After data cleaning and application of exclusion criteria, our study population consisted of 1431 patients in 21 Italian centers (10 oncological and 11 nononcological centers) (Figure 1 and Supplemental Table S1). Based on histology, 995/1431 (69.5%) tumors were benign and 436/1431 (30.5%) were malignant (115/1431, 8.0% borderline; 263/1431, 18.4% primary invasive; 58/1431, 4.1% metastatic tumors). Tumor outcome according to center is shown in Supplemental Table S1. Clinical, ultrasound, and histological characteristics of the study population are summarized in Table 1 and Supplemental Table S2. Information on CA125 was missing in 394/1431 (27.5%) patients. Missing CA125 values were less common in patients who had a malignant than benign diagnosis based on subjective assessment at the inclusion scan (Supplemental Table S3). The characteristics of our study population and those of the studies in which the RMI and IOTA models were developed are shown in Supplemental Table S4.

Risk of Malignancy Index

The overall AUROC for RMI was 0.85 (95% CI, 0.81−0.87) (Figure 2). Differences in AUROC between centers (heterogeneity) are reported in Supplemental Figure S1. At a threshold of 200, which is commonly used in European guidelines,[11−15] RMI had a sensitivity of 0.58 (95% CI, 0.52−0.63) and a specificity of 0.94 (95% CI, 0.92−0.96) (Supplemental Table S5). The relationship between the RMI score and the observed prevalence of malignancy is shown in Supplemental Figure S2. At an RMI score of 200, the observed prevalence of malignancy was 55% (95% CI, 49−61).

Simple Rules Risk Model

The overall AUROC for SRRisk was 0.91 (95% CI, 0.89−0.93) (Figure 2). The SRRisk model showed small differences in

TABLE 1
Clinical, ultrasound, and histological characteristics of the study population (n = 1431)

Parameters	Median (IQR), or n (%), range
Patient age at recruitment (y)	52 (IQR, 40–62)
	Range: 18–88
Postmenopausal	745 (52)
Gynecological symptoms during the year preceding inclusion	601 (42)
Bilateral masses	281 (20)
Presence of solid components	730 (51)
Maximum diameter of lesion (mm)	69 (IQR, 48–100)
	Range: 9–400
Largest diameter of largest solid component (mm)[a]	40 (IQR, 16–68)
	Range: 2–250
Number of papillary projections	
0	1086 (76)
1	153 (11)
2	49 (3)
3	28 (2)
>3	115 (8)
More than 10 cyst locules	160 (11)
Acoustic shadows	327 (23)
Ascites	120 (8)
CA125 results missing	394 (28)
CA125 (U/mL, if available)	19 (IQR, 10–57)
	Range: 1–12,000
Color score of intratumoral flow	
1: no blood flow	733 (51)
2: minimal blood flow	277 (19)
3: moderate blood flow	239 (17)
4: very strong blood flow	182 (13)
Histological diagnosis[b]	
Benign	995 (69)
Borderline	115 (8)
Stage I primary invasive	109 (8)
Stage II–IV primary invasive	154 (11)
Secondary metastatic	58 (4)

IQR, interquartile range.

[a] For tumors with a solid component; [b] Specific histological diagnoses are shown in Supplemental Table S2.

AUROC across centers (ie, low heterogeneity) (Supplemental Figure S3). At a risk threshold of 10% (the risk threshold recommended in an international consensus statement for referring patients to an oncology center),[6] SRRisk achieved a sensitivity of 0.93 (95% CI, 0.88–0.96) and a specificity of 0.78 (95% CI, 0.69–0.85) (Supplemental Table S6). The malignancy risk was slightly underestimated by SRRisk, but SRRisk was better calibrated than ADNEX (point estimate O:E ratio 1.04; 95% CI, 0.97–1.12) (Table 2, Figure 3).

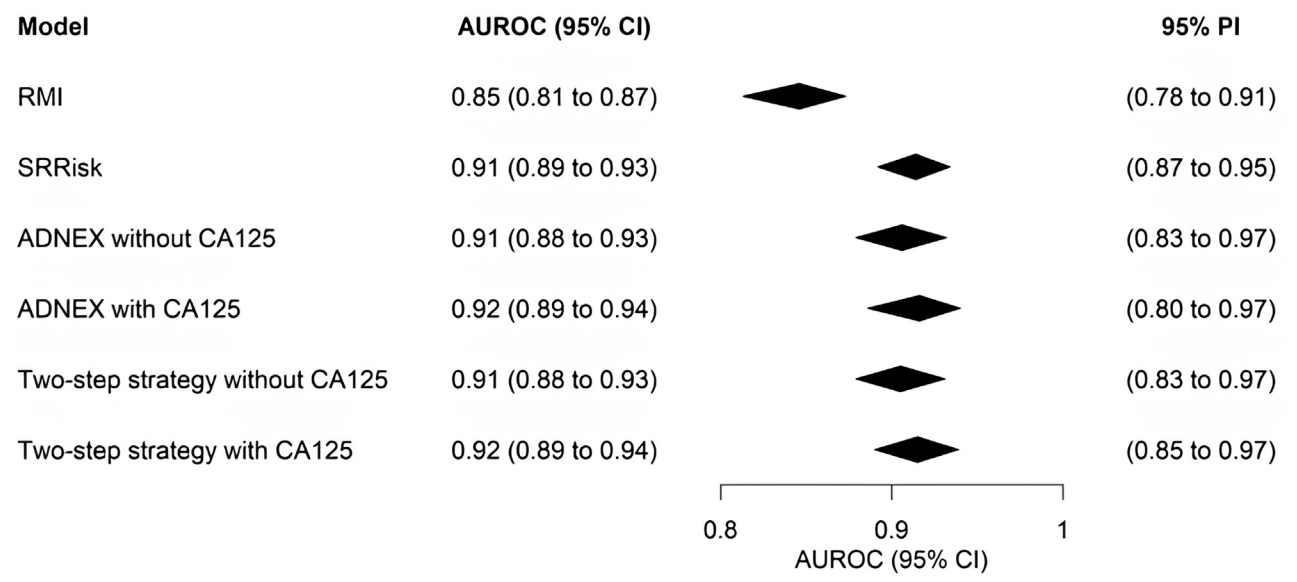
Assessment of Different NEoplasias in the adneXa model with and without Ca125

The overall AUROC for ADNEX without CA125 was 0.91 (95% CI, 0.88–0.93) and for ADNEX with CA125 it was 0.92 (95% CI, 0.89–0.94) (Figure 2). Heterogeneity across centers was slightly less for ADNEX with CA125 than for ADNEX without CA125 (Supplemental Figures S4 and S5). At a risk threshold of 10%, ADNEX without CA125 had a sensitivity of 0.94 (95% CI, 0.88–0.97) and a specificity of 0.77 (95% CI, 0.66–0.85), while ADNEX with CA125 had a sensitivity of 0.92 (95% CI, 0.87–0.95) and a specificity of 0.80 (95% CI, 0.71–0.86) (Supplemental Table S6). Malignancy risk was slightly underestimated by both ADNEX without CA125 and ADNEX with CA125 (O:E ratios 1.11 and 1.18, respectively) (Table 2, Figure 3). The ability of ADNEX to discriminate between different tumor types is shown in Table 3 and Supplemental Table S7. Adding CA125 to the ADNEX model improved discrimination between stage II to IV primary invasive malignancies and stage I primary invasive malignancies and between stage II to IV primary invasive malignancies and metastases. The calibration of ADNEX for each type of tumor is shown in Supplemental Table S8. ADNEX overestimated the likelihood of stage II to IV primary invasive ovarian malignancy but underestimated the likelihood of borderline, stage I primary invasive ovarian malignancy and a metastasis in the ovary.

Two-step strategy with and without Ca125

The overall AUROC for the 2-step strategy without CA125 was 0.91 (95% CI, 0.88–0.93) and with CA125 was 0.92 (95% CI, 0.89–0.94) (Figure 2). Heterogeneity across centers was slightly less for the 2-step strategy with CA125 than without, and the results were similar to those of ADNEX

FIGURE 2
Summary forest plot of area under the receiver operating characteristic curve (AUROC) based on meta-analysis of data from 21 centers



CI, confidence interval; *PI*, prediction interval.

(Supplemental Figures S6 and S7). The 2-step strategy showed the same classification performance as ADNEX at the 10% risk threshold, both when CA125 was included and when it was not. The ability of the 2-step strategy to discriminate between different tumor types is shown in Table 3 and Supplemental Table S7. Adding CA125 to the 2-step strategy improved discrimination between stage II to IV primary invasive malignancies and stage I primary invasive malignancies and between stage II to IV primary invasive malignancies and metastases. The calibration of the 2-step strategy for each type of tumor is shown in Supplemental Table S8. The 2-step strategy overestimated the likelihood of stage II to IV primary invasive ovarian malignancy but underestimated the likelihood of borderline, stage I primary invasive ovarian malignancy and metastasis in the ovary.

Benign descriptors, simple rules, and subjective assessment

The BDs applied to 328 of 1431 (23%) tumors, of which 325 (99%) were benign, 3 (1%; 95% CI, 1%–2%) were borderline, and none was an invasive malignancy (Supplemental Table S9).

The Simple Rules were applicable in 1244 of 1431 masses (87%). If masses with inconclusive results were classified as malignant, the sensitivity was 0.90 (95% CI, 0.85–0.93) and the specificity was 0.85 (95% CI, 0.80–0.88) (Supplemental Table S10).

Subjective assessment showed a sensitivity of 0.93 (95% CI, 0.90–0.95) and a specificity of 0.88 (95% CI, 0.84–0.91) (Supplemental Table S10).

Clinical utility

Decision curves are shown in Figure 4. RMI had the lowest clinical utility of all methods tested. At risk thresholds below 14%, RMI at cutoff 200 was less clinically useful than simply treating everyone (ie, referring all patients to an oncology center). SRRisk, ADNEX, and the 2-step strategy were clinically useful over the whole range of risk thresholds except SRRisk and ADNEX without CA125 at risk threshold 1%. At the very lowest risk thresholds (1%–5%), the 2-step strategy had higher NB than ADNEX. From risk threshold 8% and upwards, subjective assessment had the highest NB. At the 10% risk threshold, subjective

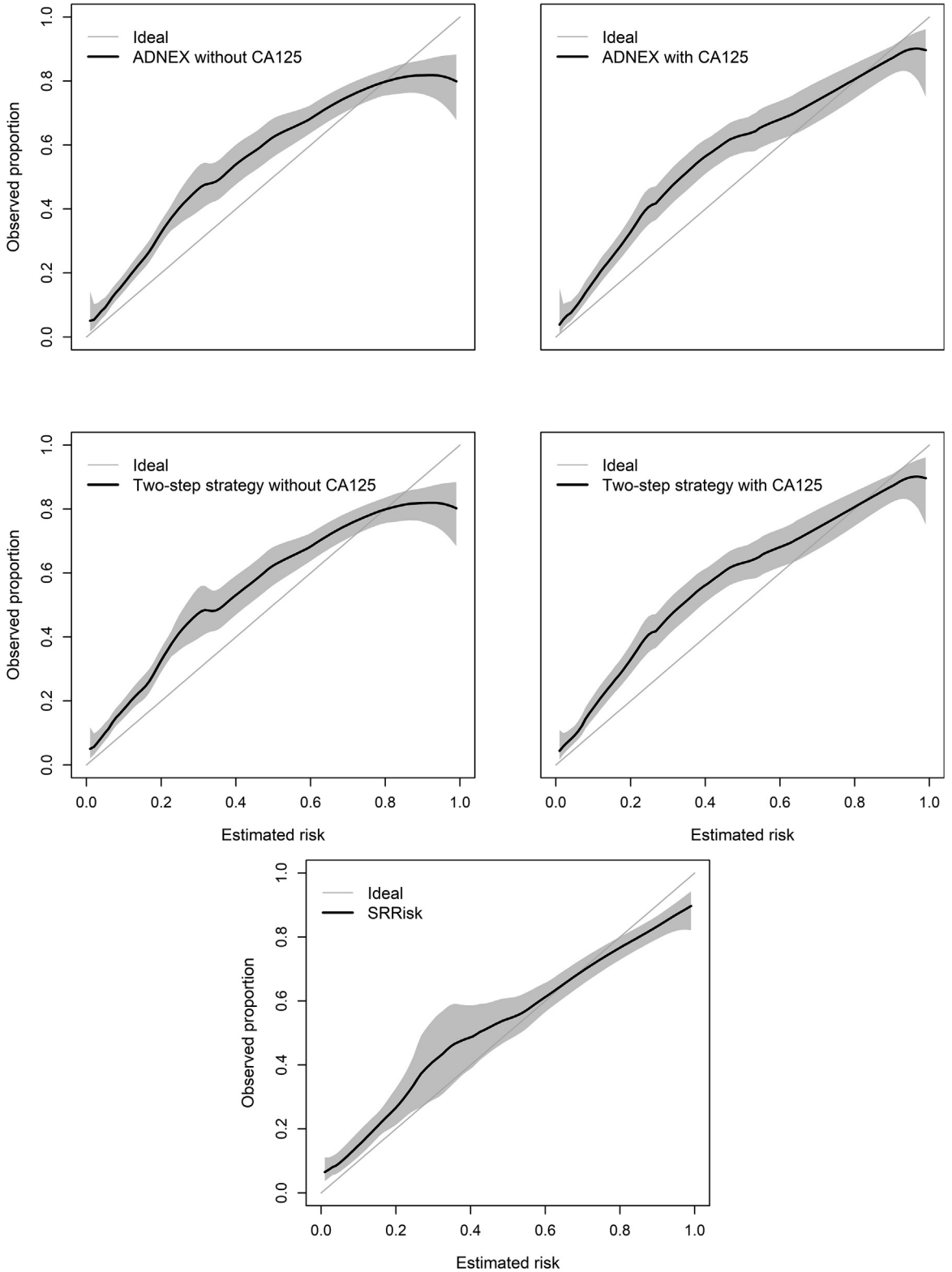
TABLE 2
Calibration in terms of observed over expected ratio based on meta-analysis of data from 21 ultrasound centers in Italy

Model	O:E ratio (95% CI)
SRRisk	1.04 (0.97; 1.12)
ADNEX without CA125	1.11 (1.04; 1.20)
ADNEX with CA125	1.18 (1.07; 1.29)
Two-step without CA125	1.13 (1.05; 1.21)
Two-step with CA125	1.20 (1.09; 1.32)

O:E ratio, observed over expected ratio. Measure of calibration in the large (mean calibration) is calculated as the observed risk of having the outcome event in the entire validation dataset divided by the average risk predicted by the model. A value >1 indicates that the model underestimates the average risk. A value <1 means that the model overestimates the average risk.

ADNEX, Assessment of Different NEoplasias in the adneXa; *CI*, confidence interval; *SRRisk*, Simple Rules Risk Model.

FIGURE 3
Flexible calibration curves using LOESS based on meta-analysis



Due to computational problems, we divided 13 centers with low sample size or low prevalence of malignancy into 4 groups: Santorso, Foggia, Treviso (group 1); Messina, Carpi, Montebelluna (group 2); Verona, Firenze, Padova, Roma (group 3); and Bari B, Asti, Bolzano (group 4). The curves were obtained with meta-analysis of center-specific curves from 8 centers and from the 4 groups. The shaded areas represent the 95% confidence band around the calibration curve.

ADNEX, Assessment of Different NEoplasias in the adneXa; *LOESS*, locally estimated scatterplot smoothing; *SRRisk*, Simple Rules Risk Model.

TABLE 3
Polytomous discrimination index (PDI) of Assessment of Different NEoplasias in the adneXa (ADNEX) and of the 2-step strategy (pooled analysis)

Model	PDI (95% CI)
ADNEX without CA125	0.49 (0.47; 0.53)
ADNEX with CA125	0.55 (0.51; 0.59)
Two-step strategy without CA125	0.49 (0.47; 0.52)
Two-step strategy with CA125	0.55 (0.51; 0.59)

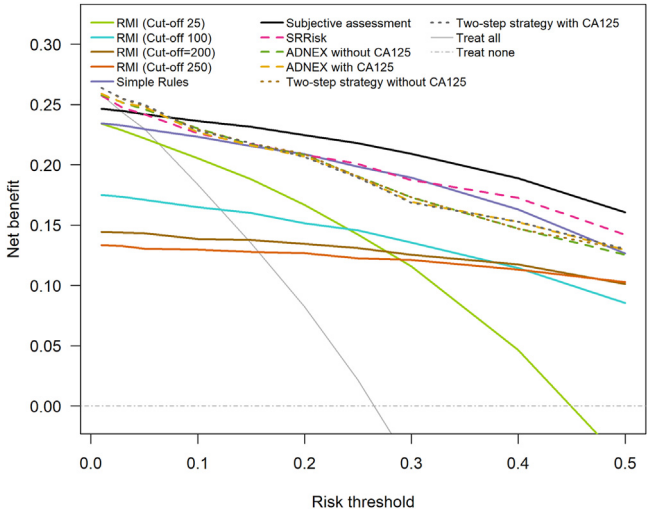
ADNEX, Assessment of Different NEoplasias in the adneXa; *CI*, confidence interval; *PDI*, Polytomous Discrimination Index.

assessment had the highest NB, the NB of ADNEX, and 2-step strategy were similar and higher than that of SSRisk and Simple Rules, while RMI at 100, 200, or 250 was not clinically useful at this threshold.

Subgroup analyses

Tumor outcome and percentage of missing CA125 values in subgroups according to menopausal status, type of center, and ultrasound examiners' level of expertise are shown in Table 4. In all subgroups, the AUROCs were higher for the IOTA models than for RMI (Figure 5). The AUROCs of the IOTA models were >0.90 (0.90–0.95) in all subgroups. They were slightly higher in premenopausal than postmenopausal patients, in nononcology than oncology centers, and for EFSUMB level 3 ultrasound examiners than for EFSUMB level 1 or 2 ultrasound examiners. The sensitivity of subjective assessment was higher for level 3 and level 2 examiners than for level 1 examiners (0.96 vs 0.92 vs 0.86), but the corresponding specificity was lower (0.84 vs 0.87 vs 0.97) (Supplemental Table S11). The sensitivity of Simple Rules (inconclusive cases classified as malignant) was higher for level 3 than for level 2 or 1 examiners (0.95 vs 0.88 vs 0.82) with the corresponding specificity being lower (0.80 vs 0.85 vs 0.92). Calibration of the IOTA models in the subgroups is shown in Supplemental Figure S8. In all subgroups, the SRRisk model was better calibrated than the ADNEX model (with and without CA125) and the 2-step strategy (with and without CA125).

FIGURE 4
Decision curves for risk models, Risk of Malignancy Index (RMI), Simple Rules, and subjective assessment based on meta-analysis of data from 21 centers



The curves show net benefit at several thresholds (exchange rates) between 1% and 50%. A model is clinically useful if it is superior to both treat all and treat none. At risk thresholds below 14%, using RMI at cutoff 200, had less clinical utility than treating everyone (ie, it was less clinically useful than referring all women with an adnexal mass to a gynecological oncology center).

Ovarian-Adnexal Reporting and Data System

The prevalence of malignancy in the 4 O-RADS risk groups is shown in Table 5 together with the agreement between the 2 risk calculation methods. Using the 2-step strategy, the prevalence of malignancy per O-RADS group was: O-RADS 2: 0.97% (95% CI, 0.4–2.6), O-RADS 3: 7.2% (95% CI, 5.0–10.3), O-RADS 4: 37.9% (95% CI, 32.4–43.8), and O-RADS 5: 84.0% (95% CI, 79.8–87.4). Using the O-RADS lexicon, the prevalence of malignancy per O-RADS group was O-RADS 2: 1.2% (95% CI, 0.5–2.9), O-RADS 3: 6.0% (95% CI, 3.6–9.6), O-RADS 4: 27.8% (95% CI, 23.6–32.5), and O-RADS 5: 83.1% (95% CI, 79.0–86.6).

Comment

Principal findings

SRRisk, ADNEX, and the IOTA 2-step strategy (with or without CA125) discriminated well between benign and malignant adnexal masses and were superior to RMI when validated on a national basis in 21 Italian centers by IOTA-certified gynecologists with different levels of ultrasound expertise. All IOTA methods had higher NB than RMI. The prevalence of malignancy in each of the 4 O-RADS risk groups closely matched the assigned malignancy risk regardless of whether the 2-step strategy or O-RADS lexicon was used.

Strengths and limitations

Our study is the first prospective national multicenter study to validate IOTA models and to validate them in the

TABLE 4
Tumor outcome and percentage of missing CA125 values for all prespecified subgroups

	Outcome			
Subgroup	N	Benign	Malignant	Missing CA125
Postmenopausal	745	444 (60)	301 (40)	176 (24)
Premenopausal	686	551 (80)	135 (20)	218 (32)
Oncology center	817	583 (64)	331 (36)	219 (27)
Other center	614	412 (80)	105 (20)	175 (28)
Level of experience of ultrasound examiners				
<500 scans	123	102 (83)	21 (17)	41 (33)
500–5000 scans	650	476 (73)	174 (27)	108 (17)
>5000 scans	658	417 (63)	241 (37)	245 (37)
EFSUMB level of ultrasound examiners				
Level 1	118	96 (81)	22 (19)	38 (32)
Level 2	884	605 (68)	279 (32)	248 (28)
Level 3	429	294 (69)	135 (31)	108 (25)

Results are shown as n (%).

EFSUMB, European Federation of Societies for Ultrasound in Medicine and Biology.

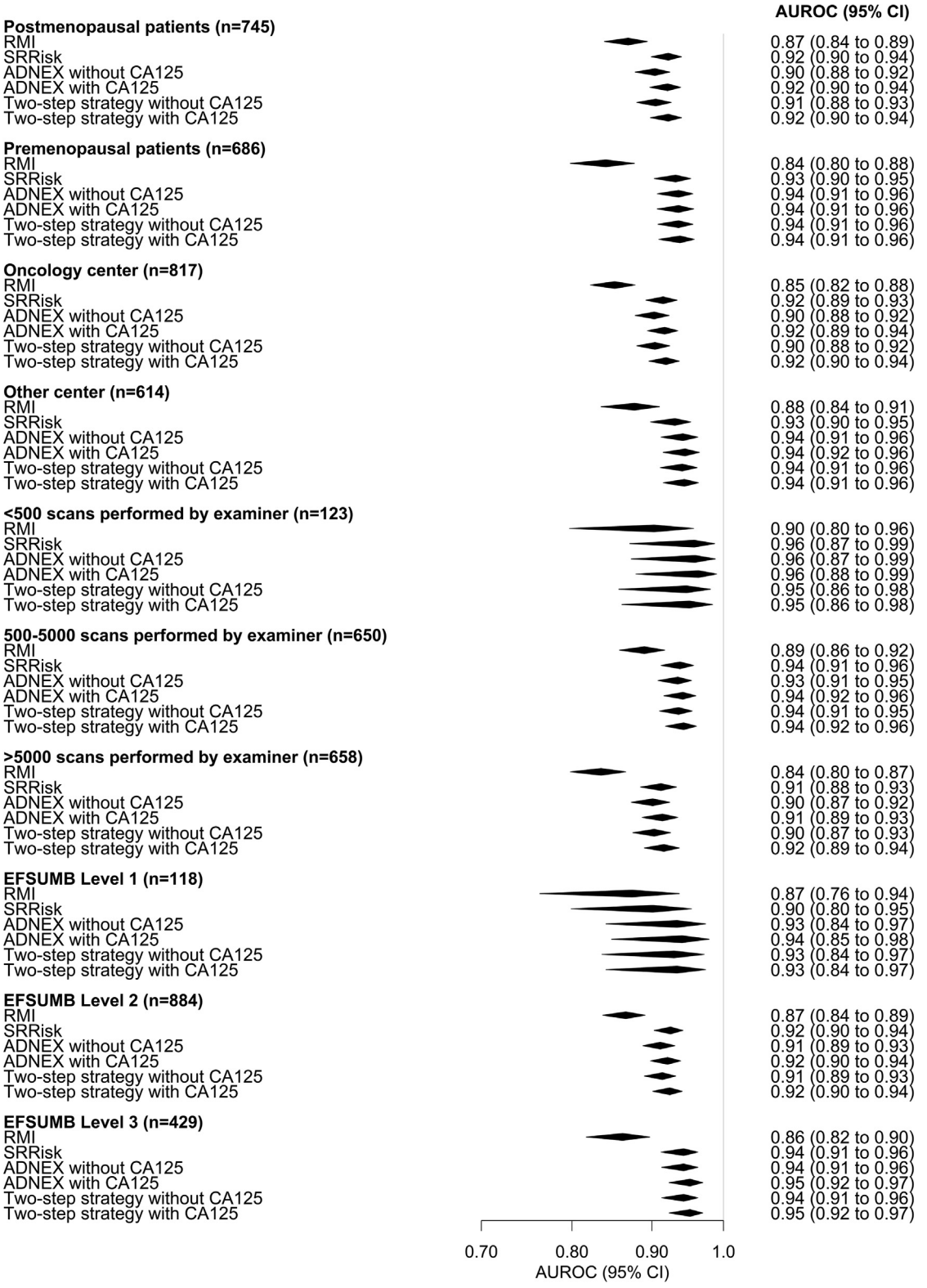
hands of ultrasound examiners with different levels of expertise. Limitations are the small number of EFSUMB level 1 examiners, the small number of centers from the south of Italy (our aim was to include centers homogenously distributed all over Italy), and that CA125 was missing in 28% of patients. We addressed the missing CA125 values using multiple imputations. Excluding participants with missing CA125 results leads to selection bias.[52–55] Multiple imputation is a recommended approach to avoid such exclusions.[55] Using histology as reference standard can be seen both as a strength and as a limitation. The strength is that using the same reference standard in all patients avoids differential verification bias. The limitation is that our results might not be applicable to all adnexal masses, which include also those managed with clinical and ultrasound follow-up. Another limitation of our study is the retrospective application of the O-RADS classification, which may introduce bias resulting in overestimation of the performance of O-RADS. However, prospective application was not feasible, as O-RADS was published in 2020, whereas the IOTA phase 6 study was designed in 2016. Despite this, our analysis of the performance of O-RADS has some value, because the ultrasound data, including detailed descriptions of the adnexal masses, were collected prospectively, allowing for a reasonably accurate retrospective application of the O-RADS lexicon. A prospective, multicenter, and ideally international study involving examiners with varying levels of ultrasound expertise and including both surgically and conservatively managed patients will be necessary to confirm or refute our findings regarding O-RADS. Finally, another limitation could be the use of different ultrasound devices, as the equipment was not standardized across centers. This may have influenced image quality and the detection of certain morphological features. On the other hand, use of different ultrasound equipment reflects clinical practice and increases the likelihood that our results are generalizable.

Comparison with other studies

The results of our study agree well with those in other validation studies.[9,17,22,25,33,56–58] The discriminative performance of ADNEX in our study (AUROC 0.92 and 0.91 for ADNEX with and without CA125, respectively) was similar to that reported in a meta-analysis including 17,007 adnexal masses examined with ultrasound in different countries and settings in 47 studies (AUROC 0.93 both for ADNEX with and without CA125).[22] It was only slightly poorer than that in a large international multicenter study conducted by the IOTA group (AUROC 0.94 both for ADNEX with and without CA125) and in a large single-center study conducted in a private center in Barcelona, Spain (AUROC, 0.95).[24,59] The discriminative performance of the 2-step strategy was lightly poorer in our study than in 2 other large studies[17,59] (AUROC 0.92 vs 0.95[17] when ADNEX with CA125 was used as second step test; AUROC 0.91 vs 0.94[17] vs 0.95[59] when ADNEX without CA125 was used as second step test). Whether the small differences in discriminative performance are explained by differences in tumor characteristics (the studies cited included also patients managed expectantly) or in ultrasound expertise is difficult to know. We found the discriminative ability and the clinical utility of ADNEX, the 2-step strategy, and SRRisk to be superior to those of

FIGURE 5

Forest plot of area under the receiver operating characteristic curve (AUROC) for prespecified subgroups (pooled data)



CI, confidence interval; *EFSUMB*, European Federation of Societies for Ultrasound in Medicine and Biology; *PI*, predicition interval.

TABLE 5
Number of patients in each Ovarian-Adnexal Reporting and Data System (O-RADS) risk group according to the risk calculation method used (O-RADS lexicon or 2-step strategy) and the observed prevalence of malignancy in each risk group

O-RADS	Lexicon					
		O-RADS 2	O-RADS 3	O-RADS 4	O-RADS 5	Observed prevalence of malignancy (95% CI)
Two-step strategy (Benign Descriptors *plus* ADNEX with CA125)	O-RADS 2	346	41	13	0	4/411[a] 0.97% (0.4; 2.6)
	O-RADS 3	52	178	140	12	28/387[b] 7.2% (5.0; 10.3)
	O-RADS 4	9	28	170	70	105/277 37.9% (32.4; 43.8)
	O-RADS 5	2	5	58	291	299/356 84.0% (79.8; 87.4)
	Observed prevalence of malignancy	5/409 1.2% (0.5; 2.9)	15/252 6.0% (3.6; 9.6)	106/381 27.8% (23.6; 32.5)	310/373 83.1% (79.0; 86.6)	

The observed prevalence of malignancy for O-RADS with the 2-step strategy is reported for all 1431 patients.

The observed prevalence of malignancy for O-RADS with the lexicon is reported for 1415 patients, because 16 of 1431 patients could not be classified based on lexicon descriptors derived from the International Ovarian Tumor Analysis (IOTA) variables.

ADNEX, Assessment of Different NEoplasias in the adneXa; *CI*, confidence interval; *O-RADS*, Ovarian-Adnexal Reporting and Data System.

[a] 11 patients classified in O-RADS 2 with the 2-step strategy could not be classified in any O-RADS category using the lexicon (this explains denominator 411); [b] Five patients classified in O-RADS 3 with the 2-step strategy could not be classified in any O-RADS category using the lexicon (this explains denominator 387).

RMI, which agrees with the results of another study.[24] Both in our study and in others, the IOTA models underestimated the risk of malignancy, the best calibrated model being SRRisk, and the models being better calibrated in postmenopausal than premenopausal patients.[17,24]

We found the sensitivity of subjective assessment to be 0.93 and the specificity to be 0.88, which is almost identical to the sensitivity and specificity of subjective assessment reported in a meta-analysis (sensitivity 0.93 and specificity 0.89).[9] The classification performance of subjective assessment is heavily dependent on the expertise of the ultrasound examiner.[60] The performance of risk calculation models should be less dependent on ultrasound skill as long as the ultrasound examiner is familiar with the definitions of the variables in the models. Nonetheless, we found some small differences in the discriminative and calibration performance of the IOTA models between examiners with different levels of expertise, with performance being slightly better in the group of EFSUMB level 3 examiners. However, it is difficult to interpret these differences, because they may be explained by a difference in tumor types between the groups.

Our findings align closely with those of a retrospective study by Timmerman et al,[35] which externally validated the ability of the IOTA 2-step strategy and O-RADS lexicon to classify adnexal masses into the O-RADS risk groups in 4905 patients. The prevalence of malignancy in the O-RADS risk groups when using the O-RADS lexicon in the study by Timmerman et al was similar to that in ours: 1% vs 1% for O-RADS 2, 4% vs 6% for O-RADS 3, 27% vs 28% for O-RADS 4, and 78% vs 83% for O-RADS 5. Both studies suffer from the limitation that the O-RADS lexicon was derived from IOTA variables, not applied prospectively.

Implications in clinical practice

The good performance of the IOTA models in our study, which includes also ultrasound examiners with limited ultrasound experience and local, regional, and university hospitals, supports that IOTA models can be widely applied in clinical practice. Our findings also support the recommendation by Landolfo et al[17] to use the 2-step strategy. The 2-step strategy had almost the same discrimination and calibration performance and almost the same clinical utility as ADNEX at risk thresholds up to 20% (and better clinical utility than ADNEX at the lowest risk thresholds) in our study, but the 2-step strategy is easier to use than ADNEX while still offering the advantage of providing an estimate of the likelihood of 4 types of malignancy.

The O-RADS lexicon also appears to be promising for clinical application. However, in our study, the O-RADS lexicon was retrospectively derived from prospectively collected IOTA variables. It will be interesting to assess how well the O-RADS lexicon performs when applied in a prospective study.

Future perspectives

Prospective studies including a very large number of ultrasound examiners with limited ultrasound experience are

needed to confirm our results. It would be important to investigate the effect of using the IOTA models in impact studies.[61] Such studies will show whether the use of IOTA models in daily practice improves decision-making and, ultimately, patient outcomes.

Conclusion

SRRisk, ADNEX, and the 2-step strategy with or without CA125 have similar and good ability to distinguish benign from malignant adnexal tumors in patients examined by either expert or nonexpert ultrasound operators in Italy, and they are all superior to RMI. Our results support the recommendation by the IOTA group to use the 2-step strategy to characterize ovarian tumors. ■

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA Cancer J Clin 2024;74:12–49.

2. Woo YL, Kyrgiou M, Bryant A, Everett T, Dickinson HO. Centralisation of services for gynaecological cancers — a Cochrane systematic review. Gynecol Oncol 2012;126:286–90.

3. Engelen MJA, Kos HE, Willemse PHB, et al. Surgery by consultant gynecologic oncologists improves survival in patients with ovarian carcinoma. Cancer 2006;106:589–98.

4. Vernooij F, Heintz APM, Witteveen PO, van der Heiden-van der Loo M, Coebergh JW, van der Graaf Y. Specialized care and survival of ovarian cancer patients in The Netherlands: nationwide cohort study. JNCI J Nat Can Ins 2008;100:399–406.

5. Froyman W, Landolfo C, De Cock B, et al. Risk of complications in patients with conservatively managed ovarian tumours (IOTA5): a 2-year interim analysis of a multicentre, prospective, cohort study. Lancet Oncol 2019;20:448–58.

6. Timmerman D, Planchamp F, Bourne T, et al. ESGO/ISUOG/IOTA/ESGE Consensus Statement on preoperative diagnosis of ovarian tumors. Ultrasound Obstet Gynecol 2021;58:148–68.

7. Valentin L, Hagen B, Tingulstad S, Eik-Nes S. Comparison of 'pattern recognition' and logistic regression models for discrimination between benign and malignant pelvic masses: a prospective cross validation. Ultrasound Obstet Gynecol 2001;18:357–65.

8. Timmerman D. The use of mathematical models to evaluate pelvic masses; can they beat an expert operator? Best Pract Res Clin Obstet Gynaecol 2004;18:91–104.

9. Meys EMJ, Kaijser J, Kruitwagen RFPM, et al. Subjective assessment versus ultrasound models to diagnose ovarian cancer: a systematic review and meta-analysis. Eur J Cancer 2016;58:17–29.

10. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. BJOG 1990;97:922–9.

11. https://kunskapsbanken.cancercentrum.se/diagnoser/aggstockscancer-epitelial/vardprogram/diagnostik/. Accessed July 7, 2025.

12. https://www.legeforeningen.no/contentassets/75de8a48892f4cdc8db45fc7804d369c/benigne-ovarialcyster_2024.pdf. Accessed July 7, 2025.

13. https://static1.squarespace.com/static/5467abcce4b056d72594db79/t/58ed3a3ddb29d654c6c.d634c/14919419526677/Cysteguideline.pdf. Accessed July 7, 2025.

14. Royal College of Obstetricians and Gynaecologists. The management of ovarian cysts in postmenopausal women. Green-top Guideline No. 34; 2016.

15. Sundar S, Agarwal R, Davenport C, et al. Risk-prediction models in postmenopausal patients with symptoms of suspected ovarian cancer in the UK (ROCkeTS): a multicentre, prospective diagnostic accuracy study. Lancet Oncol 2024;25:1371–86.

16. Ameye L, Timmerman D, Valentin L, et al. Clinically oriented three-step strategy for assessment of adnexal pathology. Ultrasound Obstet Gynecol 2012;40:582–91.

17. Landolfo C, Bourne T, Froyman W, et al. Benign descriptors and ADNEX in two-step strategy to estimate risk of malignancy in ovarian tumors: retrospective validation in IOTA5 multicenter cohort. Ultrasound Obstet Gynecol 2023;61:231–42.

18. Timmerman D, Testa AC, Bourne T, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol 2008;31:681–90.

19. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the international ovarian tumor analysis group. J Clin Oncol 2005;23:8794–801.

20. Timmerman D, Van Calster B, Testa A, et al. Predicting the risk of malignancy in adnexal masses based on the simple rules from the international ovarian tumor analysis group. Am J Obstet Gynecol 2016;214:424–37.

21. Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. BMJ 2014;349:g5920.

22. Barreñada L, Ledger A, Dhiman P, et al. ADNEX risk prediction model for diagnosis of ovarian cancer: systematic review and meta-analysis of external validation studies. BMJ Med 2024;3:e000817.

23. Andreotti RF, Timmerman D, Strachowski LM, et al. O-RADS US risk stratification and management system: a consensus guideline from the ACR ovarian-adnexal reporting and data system committee. Radiology 2020;294:168–85.

24. Van Calster B, Valentin L, Froyman W, et al. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. BMJ 2020;370:m2614.

25. Hiett AK, Sonek JD, Guy M, Reid TJ. Performance of IOTA Simple Rules, Simple Rules risk assessment, ADNEX model and O-RADS in differentiating between benign and malignant adnexal lesions in North American women. Ultrasound Obstet Gynecol 2022;59:668–76.

26. Jeong SY, Park BK, Lee YY, Kim TJ. Validation of IOTA-ADNEX model in discriminating characteristics of adnexal masses: a comparison with subjective assessment. J Clin Med 2020;9:2010.

27. Esquivel Villabona AL, Rodríguez JN, Ayala N, et al. Two-step strategy for optimizing the preoperative classification of adnexal masses in a university hospital, using international ovarian tumor analysis models. J Ultrasound Med 2022;41:471–82.

28. Rashmi N, Singh S, Begum J, Sable MN. Diagnostic performance of ultrasound-based international ovarian tumor analysis simple rules and assessment of different NEoplasias in the adneXa model for predicting malignancy in women with ovarian tumors: a prospective cohort study. Women's Health Rep 2023;4:202–10.

29. Velayo CL, Reforma KN, Sicam RVG, Diwa MH, Sy ADR, Tantengco OAG. Diagnostic performances of ultrasound-based models for predicting malignancy in patients with adnexal masses. Healthcare 2022;11:8.

30. Qian L, Du Q, Jiang M, Yuan F, Chen H, Feng W. Comparison of the diagnostic performances of ultrasound-based models for predicting malignancy in patients with adnexal masses. Front Oncol 2021;11:673722.

31. Araujo KG, Jales RM, Pereira PN, et al. Performance of the IOTA ADNEX model in preoperative discrimination of adnexal masses in a gynecological oncology center. Ultrasound Obstet Gynecol 2017;49:778–83.

32. Timmerman D, Van Calster B, Testa AC, et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. Ultrasound Obstet Gynecol 2010;36:226–34.

33. Timmerman D, Ameye L, Fischerova D, et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. BMJ 2010;341:c6839.

34. Sayasneh A, Ferrara L, De Cock B, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model: a multicentre external validation study. Br J Cancer 2016;115:542–8.

35. Timmerman S, Valentin L, Ceusters J, et al. External validation of the ovarian-adnexal reporting and data system (O-RADS) lexicon and the international ovarian tumor analysis 2-step strategy to stratify ovarian tumors into O-RADS risk groups. JAMA Oncol 2023;9:225.

36. Poonyakanok V, Tanmahasamut P, Jaishuen A, et al. Preoperative evaluation of the ADNEX model for the prediction of the ovarian cancer risk of adnexal masses at Siriraj hospital. Gynecol Obstet Invest 2021;86:132–8.

37. Giourga M, Pouliakis A, Vlastarakos P, et al. Evaluation of IOTA-ADNEX model and simple rules for identifying adnexal masses by operators with varying levels of expertise: a single-center diagnostic accuracy study. Ultrasound Int Open 2023;09:E11–7.

38. Grover SB, Patra S, Grover H, Mittal P, Khanna G. Prospective revalidation of IOTA "two-step", "alternative two-step" and "three-step" strategies for characterization of adnexal masses – An Indian study focussing the radiology context. Indian J Radiol Imaging 2020;30:304–18.

39. Debray TPA, Collins GS, Riley RD, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ 2023;380:e071018.

40. https://iotaplus.org/en/certified-members. Accessed July 7, 2025.

41. European federation of societies for ultrasound in medicine and biology. Ultraschall Med - Eur J Ultrasound 2006;27:79–95.

42. Timmerman D, Valentin L, Bourne TH, et al. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol 2000;16:500–5.

43. Prat J. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. Int J Gynecol Obstet 2014;124:1–5.

44. Installé AJ, Van den Bosch T, De Moor B, Timmerman D. Clinical data miner: an electronic case report form system with integrated data preprocessing and machine-learning libraries supporting clinical diagnostic model research. JMIR Med Inform 2014;2:e28.

45. Riley RD, Debray TPA, Collins GS, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med 2021;40:4230–51.

46. Debray TPA, Damen JAAG, Snell KIE, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.

47. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76.

48. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 2006;26:565–74.

49. Lin L, Chu H. Meta-analysis of proportions using generalized linear mixed models. Epidemiology 2020;31:713–7.

50. Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW. Extending the *c* -statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. Stat Med 2012;31:2610–26.

51. Van Calster B, Vergouwe Y, Looman CWN, Van Belle V, Timmerman D, Steyerberg EW. Assessing the discriminative ability of risk models for more than two outcome categories. Eur J Epidemiol 2012;27:761–70.

52. Moons KGM, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1–73.

53. Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.

54. de Groot JAH, Bossuyt PMM, Reitsma JB, et al. Verification problems in diagnostic accuracy studies: consequences and solutions. BMJ 2011;343:d4770.

55. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019;170:W1–33.

56. Alcázar JL, Pascual MA, Graupera B, et al. External validation of IOTA simple descriptors and simple rules for classifying adnexal masses. Ultrasound Obstet Gynecol 2016;48:397–402.

57. Kaijser J, Sayasneh A, Van Hoorde K, et al. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. Hum Reprod Update 2014;20:449–62.

58. Westwood M, Ramaekers B, Lang S, et al. Risk scores to guide referral decisions for people with suspected ovarian cancer in secondary care: a systematic review and cost-effectiveness analysis. Health Technol Assess (Rockv) 2018;22:1–264.

59. Pascual MA, Vancraeynest L, Timmerman S, et al. Validation of ADNEX and IOTA two-step strategy and estimation of risk of complications during follow-up of adnexal masses in low-risk population. Ultrasound Obstet Gynecol 2024;64:395–404.

60. Timmerman D, Schwärzler P, Collins WP, et al. Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. Ultrasound Obstet Gynecol 1999;13:11–6.

61. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ 2009;338:b606.

Author and article information

From the UniCamillus, International Medical University, Rome, Italy (Moro); Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy (Moro, Momi, Mascilini, Ciccarone, Pozzati, Fagotti and Testa); Division of Obstetrics and Gynecology, University of Brescia, Brescia, Italy (Momi); Department of Development and Regeneration, KU Leuven, Leuven, Belgium (Ledger, Barreñada, Ceusters, Froyman, Calster and Timmerman); Ospedale Filippo del Ponte Ostetricia e Ginecologia, Lombardia, Varese, Italy (Sturla); Fondazione Poliambulanza Istituto Ospedaliero, Lombardia, Brescia, Italy (Mor); Azienda Ospedaliero Universitaria Pisana, Toscana, Pisa, Italy (Fornari); Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Leuven, Belgium (Froyman and Timmerman); Imperial College Healthcare NHS Trust, London, UK (Bourne); Dipartimento Universitario Scienze della Vita e Sanità Pubblica, Università Cattolica del Sacro Cuore, Rome, Italy (Fagotti and Testa); Skåne University Hospital, Malmö, Sweden (Valentin); and Department of Clinical Sciences Malmö, Lund University, Malmö, Sweden (Valentin).

*IOTA 6 Collaborators: Valentina Bertoldo[a], Fabio Ghezzi[b], Antonella Vimercati[c], Saverio Tateo[d], Marianna Roccio[e], Rosalba Giacchello[f], Roberta Granese[g], Daniela Garbin[h], Tiziana De Grandis[i], Federica Piccini[j], Patrizia Favaro[k], Olga Petruccelli[l], Anila Kardhashi[m], Ilaria Pezzani[n], Patrizia Ragno[o], Laura Falchi[p], Bruna Anna Virgilio[q], Erika Fruscella[r], Tiziana Tagliaferri[s], Annibale Mazzocco[t]

[a]Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy; [b]University of Insubria, Ospedale di Circolo Fondazione Macchi, Varese, Italy; [c]Azienda Ospedaliero-Universitaria Consorziale Policlinico di Bari, Puglia, Bari, Italy; [d]OspedaleSanta Chiara di Trento, Trentino-Alto Adige, Trento, Italy; [e]Department of Obstetrics and Gynecology, Fondazione IRCCS Policlinico San Matteo and University of Pavia, Pavia, Italy; [f]Ospedale Regina Montis Regalis Mondovi, Mondovi, Italy; [g]Obstetrics and Gynecology Unit, Department of Human Pathology of Adult and Childhood "G. Barresi", University Hospital "G. Martino", Messina, Italy; [h]Ospedale di Santorso; [i]Candiolo Cancer Institute, FPO-IRCCS-Candiolo, Turin, Italy; [j]"Ramazzini" di Carpi; [k]Ostetricia e Ginecologia, Ospedale "Orlandi di Bussolengo" Varese, Italy; [l]Azienda Ospedaliera Universitaria O.O.R.R. Foggia; [m]Istituto Tumori, IRCCS, "Giovanni Paolo II; [n]Presidio Ospedaliero di Treviso; [o]ASL AT - Asti; [p]Azienda USL Toscana Centro; [q]Ospedale di Abano Terme; [r]Ospedale Santo Spirito - Roma; [s]Ospedale San Maurizio di Bolzano; [t]Presidio Ospedaliero di Montebelluna, Veneto, Italy.

Received April 14, 2025; revised July 8, 2025; accepted July 8, 2025.

L.V. and A.C.T. contributed equally to this work.

The authors report no conflict of interest.

Corresponding author: Francesca Moro, PhD. morofrancy@gmail.com