# External validation of ultrasound-based models for discrimination between benign and malignant adnexal masses in Italy: the prospective multicenter IOTA phase 6 study

Francesca Moro[1,2], Marina Momi[2], Valentina Bertoldo[2], Ashleigh Ledger[3], Lasai Barreñada[3], Jolien Ceusters[3], Davide Sturla[4], Fabio Ghezzi[4], Elisa Mor[5], Letizia Fornari[6], Antonella Vimercati[7], Saverio Tateo[8], Marianna Roccio[9], Rosalba Giacchello[10], Roberta Granese[11], Daniela Garbin[12], Tiziana De Grandis[13], Federica Piccini[14], Patrizia Favaro[15], Olga Petruccelli[16], Anila Kardhashi[17], Ilaria Pezzani[18], Patrizia Ragno[19], Laura Falchi[20], Bruna Anna Virgilio[21], Erika Fruscella[22], Tiziana Tagliaferri[23], Annibale Mazzocco[24], Floriana Mascilini[2], Francesca Ciccarone[2], Federica Pozzati[2], Wouter Froyman[3,25], Ben Van Calster[3], Tom Bourne[26], Dirk Timmerman[3,25], Giovanni Scambia[2], *Lil Valentin[27, 28], *Antonia Carla Testa[2]

[1] UniCamillus, International Medical University, Rome, Italy
[2] Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy
[3] Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[4] Ospedale Filippo del Ponte Ostetricia e Ginecologia, Varese, Lombardia, Italy
[5] Fondazione Poliambulanza Istituto Ospedaliero, Brescia, Lombardia, Italy
[6] Azienda Ospedaliero Universitaria Pisana, Pisa, Toscana, Italy
[7] Azienda Ospedaliero-Universitaria Consorziale Policlinico di Bari, Bari, Puglia, Italy
[8] Ospedale Santa Chiara di Trento, Trento, Trentino-Alto Adige, Italy
[9] Fondazione IRCCS Policlinico San Matteo, Pavia, Lombardia, Italy
[10] ASL CN1, Cuneo, Piemonte, Italy
[11] Azienda Ospedaliera Universitaria Policlinico “G. MARTINO”
[12] Ospedale di Santorso
[13] IRCCS Candiolo
[14] "Ramazzini" di Carpi
[15] Ospedale di Bussolengo
[16] Azienda Ospedaliera Universitaria O.O.R.R. Foggia
[17] Istituto Tumori, IRCCS, "Giovanni Paolo II
[18] Presidio Ospedaliero di Treviso
[19] ASL AT - Asti
[20] Azienda USL Toscana Centro
[21] Ospedale di Abano Terme
[22] Ospedale Santo Spirito - Roma
[23] Ospedale San Maurizio di Bolzano
[24] Presidio Ospedaliero di Montebelluna, Veneto, Italy
[25] Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Leuven, Belgium
[26] Imperial College Healthcare NHS Trust, London, UK
[27] Skåne University Hospital, Malmö, Sweden
[28] Department of Clinical Sciences Malmö, Lund University, Sweden

*The last two authors contributed equally to the work.

**Corresponding authors:**

Francesca Moro

UniCamillus, International Medical University, Rome, Italy

Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy

E-mail: morofrancy@gmail.com; francesca.moro@guest.policlinicogemelli.it;

Phone number: +39-3493905802

**Keywords:**

Ovarian neoplasm, Ultrasonography,

## Abstract

**Objective** To prospectively validate the performance of the Risk of Malignancy Index (RMI), International Ovarian Tumor Analysis (IOTA) Simple Rules Risk Model (SRRisk), IOTA Assessment of Different NEoplasias in the adneXa (ADNEX) and the IOTA two-step strategy in different types of ultrasound centers in Italy.

**Methods** This is a multicenter prospective observational study including regional referral centers and district hospitals in Italy. Consecutive patients with an adnexal mass examined with ultrasound by an IOTA certified ultrasound examiner with different levels of experience were included, provided they underwent surgery < 180 days after the inclusion scan. Ultrasound examination was performed transvaginally or transrectally and/or transabdominally based on the characteristics of the women and masses. Reference standard was the histology of the adnexal mass following surgical removal. Discrimination (area under receiver operating characteristic curve, AUROC), calibration, and clinical utility were assessed to illustrate the diagnostic performance of the methods. The performance of the models was also evaluated in predefined subgroups based on menopausal status, type of center (oncology vs non-oncology) and ultrasound examiner´s experience: [<500 scans performed, 500-5000 scans performed, >5000 scans performed; European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) Level 1, Level 2, Level 3].

**Results** 1567 patients were recruited between May 2017 and March 2020 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 non-oncological). 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). For all IOTA models (SRRisk, ADNEX with and without CA125, two step strategy with and without CA125), the AUROC was between 0.91 (95% CI 0.88-0.93) and 0.92 (0.89-0.94). The AUROC was 0.85 (0.81-0.87) for RMI. The malignancy risk was slightly underestimated by all IOTA models, but

least so by SRRisk . All IOTA models had higher net benefit than RMI at risk thresholds from 1% to 50%. AUROC was ≥0.90 for all IOTA models in all subgroups, while it ranged from 0.84 to 0.90 for RMI.

**Conclusions** SRRisk, ADNEX and the two step strategy with or without CA125 had similar and good ability to distinguish benign from malignant adnexal tumours in patients examined by either expert or non-expert ultrasound operators in Italy. Their discriminative performance and clinical utility was superior to that of RMI.

## Introduction

Ovarian cancer is the leading cause of death in women diagnosed with gynecological cancers.[1] Most ovarian cancers are diagnosed at an advanced stage and require treatment in high volume centers by doctors with expertise in gynecological oncological surgery 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 optimal 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 has developed several ultrasound based methods that can be used to discriminate between benign and malignant adnexal masses: the Benign Descriptors,[16,17,] the Simple Rules, [18] four mathematical models to calculate the individual risk of malignancy in an adnexal mass (logistic regression model 1, LR1, logistic regression model 2, LR2, the Simple Rules Risk model, SRRisk, and Assessment of Different NEoplasias in the adneXa - ADNEX). [19-21] ADNEX is a multinomial regression model that calculates the probability of five outcome categories (benign, borderline, stage I primary invasive ovarian malignancy, stage II-IV primary invasive ovarian malignancy, and metastasis from another primary tumor). [21] ADNEX can be used with or without CA125 as predictor. [21] The IOTA group now recommends the IOTA two-step strategy, which means that first the Benign Descriptors are applied, and if these do not apply, ADNEX is used. [17]

The diagnostic performance of the IOTA methods and of RMI has been validated in prospective and retrospective studies, but most validation studies tested the performance in the hands of

experienced ultrasound examiners.[17,22–32] No prospective study included examiners with little ultrasound experience and few included examiners with different levels of experience.[33–35]

The primary aim of this study is to prospectively validate the diagnostic performance of RMI, SRRisk, ADNEX and the IOTA two-step strategy in different types of ultrasound centers in Italy both overall and in relevant subgroups. The secondary aims are to explore the multinomial discrimination performance of ADNEX and the two step strategy, to validate the ability of the Benign Descriptors to correctly classify an adnexal mass as benign, and to validate the classification performance of the Simple Rules and of subjective assessment overall and in the subgroups based on level of ultrasound experience.

## 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. The study was conducted in accordance with the TRIPOD cluster guidelines.

### *Patients*

Consecutive patients with a known or suspected adnexal mass examined with ultrasound by an IOTA certified ultrasound examiner[36] 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 collected 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

151 performed, surgery performed more than 180 days after the ultrasound examination, and denial or
152 withdrawal of informed consent.
153
154 *Data collection*
155 Information on age, parity, menopausal status and indication for the ultrasound examination was
156 prospectively collected, as well as information on type of hospital (private practice, local public
157 hospital, regional public hospital, or university hospital), type of center (oncological vs non-
158 oncological), and type of ultrasound center (general gynecologic outpatient clinic or specialized
159 ultrasound center). An oncological center was defined as a tertiary referral center with a dedicated
160 gynecological oncology unit. Information on the ultrasound system used, ultrasound examiner´s
161 name and level of experience was also recorded. The level of ultrasound experience was based on
162 the number of gynecological scans in non-pregnant women that the examiner had performed at the
163 start of the study. Low experience was defined as <500 scans, intermediate experience as 500-5000
164 scans, and high experience as >5000 scans. We also recorded the level of experience according to
165 the European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB level 1, 2
166 or 3 [37]), and the number of ovarian masses that the operator had examined with ultrasound during
167 the preceeding year (classified as <50, i.e. < 1 per week; 50-200, i.e. up to 4 per week; >200, i.e. > 4
168 per week).
169
170 *Ultrasound examination*
171 A standardized transvaginal (or transrectal if vaginal was not possible) ultrasound examination
172 including color or power Doppler ultrasound examination was performed, supplemented with
173 transabdominal ultrasound if transvaginal or transrectal ultrasound examination was not sufficient.
174 The IOTA examination and measurement technique were used, and the ultrasound findings were
175 described using the IOTA terminology.[38] Information on all the variables required for the IOTA
176 Benign Descriptors, Simple Rules, SRRisk, ADNEX and RMI 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) as well as the specific diagnosis suggested by the ultrasound examiner and chosen from a list of pre-defined 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 (i.e. those reported in the clinical ultrasound report), and results of other imaging modalities (e.g. computer tomography or magnetic resonance imaging), tumor markers, and patient´s preference.

*Reference standard*

Reference standard was the histology of the adnexal mass following surgical removal within 180 days after the ultrasound examination by laparotomy or laparoscopy as considered appropriate by the surgeon. Borderline tumors were classified as malignant. The histology of the surgically removed tumor was determined at each local center. Central pathology review was not performed, because we found little differences between local and central pathology reports in a previous IOTA study.[19] 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 (FIGO). [39]

*Data cleaning*

Data collection was done through the web-based clinical data miner (CDM) software.[40] Patients automatically received a unique identifier upon enrolment. We encrypted all data communication to ensure data security. A team of statisticians and ultrasound examiners performed data cleaning. Data cleaning included sending queries to participating centers to retrieve missing information or to correct inconsistencies.

*Prediction models*

We assessed the diagnostic performance of subjective assessment, RMI, Benign Descriptors, Simple rules, SRRisk, ADNEX with and without CA125, and the two-step strategy with and without CA125. Predictions were based on information obtained at the inclusion scan and so are blinded to the outcome. The results of the models were calculated after closing the study and were not used to guide patient managament.

*RMI* does not give an estimated risk but a non-negative integer (0 or higher), with higher scores suggesting a higher likelihood of malignancy. It includes clinical (CA125 and menopausal status) and ultrasound variables (multilocular cyst, solid areas, bilateral lesions, ascites, abdominal metastases).[10]

There are *four Benign Descriptors*, which classify tumors as benign: 1) a unilocular cyst with ground glass echogenicity and largest diameter <10 cm in a premenopausal woman is suggestive of endometrioma (BD 1); 2) a unilocular cyst with mixed echogenicity, acoustic shadows and largest diameter <10 cm in a premenopausal woman is suggestive of benign teratoma (BD 2); 3) a unilocular cyst with anechoic cyst fluid, smooth internal walls and largest diameter <10 cm in a pre- or post-menopausal woman is suggestive of a simple cyst or cystadenoma (BD3); 4) all other unilocular cysts with smooth internal walls and largest diameter <10 cm in a pre- or post-menopausal woman (BD 4) are suggestive of a benign cyst. [17]

*Simple Rules* classify tumors as benign, inconclusive, or malignant based on the presence of five benign ultrasound features (unilocular cyst, smooth multilocular cyst with largest diameter <100

mm, acoustic shadows, solid component(s) are present but largest diameter is <7 mm, no vascularization on color Doppler) and five malignant ultrasound features (irregular solid tumor, irregular multilocular solid tumor with largest diameter ≥100 mm, at least 4 papillary projections, presence of ascites, very strong vascularization on color Doppler). The classification is inconclusive if none of the ten features is present, or if both benign and malignant features are present. [18] We added inconclusive tumors to those predicted to be malignant, resulting in a binary classifier (benign or malignant).

*SRRisk* is a logistic regression model that calculates the risk that an adnexal tumor is malignant based on type of center (oncology center vs other center) and on the ten binary ultrasound features used in the Simple Rules. [20]

*ADNEX* is a multinomial regression model that calculates the probability of five outcome categories: benign, borderline, stage I primary invasive ovarian malignancy, stage II-IV primary invasive ovarian malignancy, and metastasis from another primary tumor. [21] One minus the probability of a benign tumor equals the estimated risk of malignancy. ADNEX includes three clinical variables (age, type of center, i.e. oncology vs non-oncology center, serum level of CA125) and six ultrasound variables (maximum diameter of the lesion in mm, proportion of solid tissue calculated as the maximum diameter of the largest solid component in mm divided by the maximum diameter of the lesion in mm, presence of more than 10 cyst locules, number of papillary projections, acoustic shadows, ascites). The variable CA125 is optional, but for multinomial discrimination ADNEX works better with than without CA125. [22]

The *two-step strategy* uses the Benign Descriptors as a first step. If a Benign Descriptor applies, the mass is classified as benign, if not, ADNEX with or without CA125 is used to estimate the risk of malignancy. [17] Information about how we estimated the risk of malignancy when a Benign Descriptor applied is found in Appendix 2.

*Statistical analysis and sample size*

We followed a prespecified statistical analysis plan. The statistical analyses were performed with R version 4.1.2. The adequacy of the sample size is discussed in Appendix 3. We report all performance measures with 95% confidence intervals (CI).

Despite it being strongly recommended to collect blood samples for measurement of serum CA125 in all patients, CA125 results were missing in some patients. Missing CA125 values were imputed. We performed multiple imputation using the method of fully conditional specification, generating 100 imputations, leading to 100 completed datasets. To estimate the missing values of CA125, we used predictive mean matching regression using the outcome and variables that are probably related to either the level of CA125 itself, or to the unavailability of CA125. The multiple imputation procedure is described in Appendix 4. To calculate the predictions for each model, we used the formula present in the original paper.

Results are presented as absolute frequency (percentage) for nominal variables and as median, interquartile range (IQR) and range (min-max) for continuous variables as appropriate. We report the percentage of tumors to which a Benign Descriptor applied and the outcome of masses to which a Benign Descriptor 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, IOTA two-step strategy) and used meta-analysis to obtain the overall AUROC per model. The heterogeneity between centers was assessed by calculating 95% prediction intervals (PI). The meta-analysis procedure is described in Appendix 5.

We 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. An O:E higher than 1 indicates that the model underestimates the risk of malignancy, and an O:E lower than 1 indicates that the model overestimates the risk of malignancy. The ideal value is 1. [41,42] We also constructed flexible calibration curves using loess. [43] We calculated center-specific O:E and calibration curves and combined them using meta-analysis. For the meta-analysis of calibration

curves, we combined 13 centers with small sample size or low prevalence of malignancy into four groups to avoid computational problems. The four groups were: Santorso, Foggia, Treviso (group 1); Messina, Carpi, Montebelluna (group 2); Verona, Firenze, Padova, Rome (group 3); and Bari B, Asti, Bolzano (group 4). The meta-analysis procedure is described in Appendix 5.

Clinical utility to decide which patients to refer for specialized oncological care was estimated using decision curve analysis for risk thresholds between 1% and 50%.[44] Net benefit is a measure of clinical utility. To know if a model is clinically useful, we compare it to treat all and treat none (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. Because RMI does not provide risk estimates, for RMI we computed clinical utility at the following fixed RMI scores: 200 (a threshold often used clinically and recommened in several national guidelines[11-15]), 250, 100 and 25. We show overall decision curves calculated using meta-analysis of center-specific decision curves. The meta-analysis procedure is described in Appendix 5.

We calculated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for subjective assessment and the Simple Rules (inconclusive cases classified as malignant) and for the risk models at risk of malignancy cut-offs of 1%, 3%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, and 50%. For RMI, we report classification performance for cut-offs 25, 100, 200, and 250. We calculated center-specific sensitivity and specificity and combined them using meta-analysis.[45] Centers with no true positive (TP) and no false positive (FP) test results at a specific threshold were excluded from the meta-analysis of PPV for that threshold. Centers with no true negative (TN) and no false negative (FN) at a specific threshold were excluded from the meta-analysis of NPV for that threshold.

To estimate the multinomial performance of ADNEX and the two-step strategy, we computed the Polytomous Discrimination Index (PDI) and calculated the AUROC for each pair of outcome categories using the conditional risk method.[46-47] To evaluate calibration, we computed O:E ratios per category. Due to the small numbers in most centers, we anticipated computational problems

when attempting to perform meta-analysis for multinomial performance. Therefore, we used the pooled dataset to estimate the multinomial performance of ADNEX and the two-step strategy.

*Subgroup analyses*

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

## 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 non-oncological centers) (Figure 1 and Supplementary 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 Supplementary Table S1. Clinical, ultrasound and histological characteristics of the study population are summarized in Table 1. Specific histological diagnoses are shown in Supplementary Table S2. The median age of the patients was 52 years (IQR 40-62, range 18 to 88), and 745 patients (52%) were postmenopausal. The median of the maximum diameter of the lesion was 69 (IQR 48-100, range 9 - 400) mm, 281 (20%) patients had bilateral masses, and 120 (8%) patients had ascites. CA125 was missing in 28% (394/1431) of patients. The characteristics of our study population and those of the studies in which the RMI and IOTA models were developed are shown in Supplementary Table S3. Patients in our study population were more frequently postmenopausal than those in the development sets of the IOTA models, the Ca125 values (when available) were lower, while acoustic shadowing and absent color Doppler signals were more common. Multilocular cysts were less common than in the development data set of the

Simple Rules, and Ca125 values were higher than in the development data set of RMI. The distribution of tumor outcome (i.e., prevalence of borderline, stage I primary invasive, stage II-IV primary invasive, metastatic tumors) in our study population is reasonably similar to that in the studies in which the IOTA models were developed.

The BDs applied to 328/1431 (23%) tumors, of which 325 (99%) were benign, 3/328 (1%, 95% CI 1-2) were borderline, and none was an invasive malignany (Supplementary Table S4).

For all IOTA models (ADNEX with and without CA125, SSRisk, two step strategy with and without CA125), the overall AUROC was $\geq$0.91, while that of RMI was 0.85 (Figure 2). Differences in AUROC between centers (heterogeneity) were smallest for SRRisk (Figure 2, Supplementary Figure S1-S6).

At a risk threshold of 10% (the risk threshold recommended in an international consensus statement for referring patients to an oncology center [6]), all IOTA models (SRRisk, ADNEX with and without Ca125 and the two-step strategy with and without Ca125) had sensitivity >0.9 with specificity ranging from 0.77 to 0.80. ADNEX and the two step strategy had the same classification performance at risk threshold 10%: sensitivity 0.92 and specificity 0.80 when CA125 was included as a predictor, sensitivity 0.94 and specificity 0.77 when CA125 was not included as a predictor (Supplementary Table S5). At a threshold of 200, RMI had sensitivity 0.58 and specificity 0.94 (Supplementary Table S6). The Simple Rules were applicable in 1244/1431 (87%) tumors and had sensitivity 0.90 and specificity 0.85 (inconclusive cases classified as malignant). Subjective assessment had sensitivity 0.93 and specificity 0.88 (Supplementary Table S7).

The malignancy risk was slightly underestimated by all IOTA models (O:E ratios 1.04 -1.20) (Table 2). SRRisk was slightly better calibrated than other IOTA models (point estimate O:E ratio 1.04; 95% CI 0.97-1.12) (Table 2, Figure 3). The relation between the RMI score and the observed prevalence of malignancy is shown in supplementary Figure S7. At RMI score 200, the observed prevalence of malignancy was 55% (95% CI 49-61).

Decision curves are shown in Figure 4. Irrespective of which RMI cutoff was used, the IOTA methods had higher net benefit than RMI. At risk thresholds lower than 15%, ADNEX and the two-step strategy had the highest net benefit, at risk thresholds above 20% SRRisk and subjective assessment had highest net benefit. RMI at cutoff 200 was less clinically useful than simply treating everyone (referring everyone to an oncology center) at risk thresholds below 15%.

The ability of ADNEX and the two-step strategy to discriminate between different tumor types is shown in Table 3 and Supplementary Table S8. The PDI ranged from 0.44 to 0.55 for all four models. ADNEX and the two-step strategy had the same ability to discriminate between the five tumor types. Performance was poorest for discrimination between Stage I primary invasive tumors and metastases (AUROC 0.70 for all four models) and between Stage II-IV primary invasive tumors and metastases (AUROC 0.59 for models without Ca125 and 0.74 for models with CA125). AUROCs ranged from 0.93 to 0.98 for distinguishing benign tumors from invasive malignancies (primary or metastastes). Calibration of the estimated risks for the five tumor types was similar for ADNEX and the two-step strategy irrespective of whether or not CA125 was included as a predictor (Supplementary Table S9). All four models underestimated the risk of borderline, stage I primary invasive, and secondary metastatic tumors (O:E >1) but overestimated the likelihood of a stage II-IV ovarian malignancy and benign tumor (O:E<1).

*Subgroup analyses*

Tumor outcome and percentage of missing CA125 values in subgroups according to menopausal status, type of center and ultrasound examiners´ level of experience 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 $\geq$0.90 in all subgroups (0.90 – 0.95). They were slightly higher in pre- than post-menopausal patients, in non-oncology 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) (Supplementary Table S10). The sensitivity of Simple Rules (inconclusive cases classified as malignant) was higher for level 3 than level 2 and 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 Supplementary Figure S8. In all subgroups, the malignancy risk tended to be underestimated (point estimate for O:E ratio >1) by the IOTA models, with less underestimation of malignancy risk in postmenopausal than premenopausal patients and in oncology than non-oncology centers. In all subgroups, SRRisk had the least understimation. The 95% CIs for O:E ratios were very wide for EFSUMB level 1 examiners and for examiners that had performed <500 scans at study start due to the small sample size for these groups.

## Discussion

### *Principal findings*

SRRisk, ADNEX and the IOTA two-step strategy (with or without CA125) discriminated well between being and malignant adnexal masses and were superior to RMI when validated on a national basis in 21 Italian centers by ultrasound examiners with different levels of ultrasound experience. All IOTA methods had higher net benefit than RMI. SSRisk, ADNEX and the two step-strategy (with or without CA125) had the highest net benefit at risk thesholds below 15%.

### *Strengths and limitations*

Our study is the first prospective national multicenter study to validate IOTA models and to validate them in the hands of ultrasound examiners with different levels of experience. We assessed the diagnostic performance both in terms of discrimination, calibration and clinical utility. Limitations are the small number of EFSUMB level 1 examiners and examiners that had perfomed <500 scans

409 at the start of the study, the small number of centers from the south of Italy (our aim was to include
410 centers homogenously distributed all over Italy), and that CA125 was missing in a substantial
411 proportion of patients (28%). We adressed the missing CA125 values using multiple imputation.
412 Using histology as reference standard can be seen both as a strength and as a limitation. The
413 strength is that using the same reference standard in all patients avoids differential verification
414 bias.The limitation is that our results might not be applicable to all adnexal masses, which include
415 also those managed with clinical and ultrasound follow-up.
416
417 *Comparison with other studies*
418 The results of our study agree well with those in other validation studies.[9,17,23,31,48-52] The
419 discriminative performance of ADNEX in our study (AUROC 0.92 and 0.91 for ADNEX with and
420 without CA125, respectively) was similar to that reported in a meta-analysis including 17007
421 adnexal masses examined with ultrasound in different countries and settings in 47 studies (AUROC
422 0.93 both for ADNEX with and without CA125). [48] It was only slightly poorer than that in a large
423 international multicenter study conducted by the IOTA group (AUROC 0.94 both for ADNEX with
424 and without CA125) and in a large single center study conducted in a private center in Barcelona,
425 Spain (AUROC 0.95). [22,52] The discriminative performance of the two-step stragey was also
426 slightly poorer than in two other large studies [17,52] (AUROC 0.92 vs 0.95 when ADNEX with
427 CA125 was used as second step test; AUROC 0.91 vs 0.94 vs. 0.95 when ADNEX without CA125
428 was used as second step test). Whether the small differences in discriminative performance are
429 explained by differences in tumor characteristics (the studies cited included also patients managed
430 expectantly) or in ultrasound expertise is difficult to know. We found the discriminative ability and
431 the clinical utility of ADNEX, the two-step strategy and SRRisk to be superior to those of RMI,
432 which agrees with the results of other studies. [17,22] Both in our study and in others, the IOTA
433 models underestimated the risk of malignancy, the best calibrated model being SRRisk, and the
434 models being better calibrated in postmenopausal than premenopusal patients. [17,22] The Benign

Descriptors were applicable in a lower proportion of patients in our study than in those by Landolfo et al and Pascual et al (23% vs 37% vs 77%). [17,52] This is explained by the the other two studies including also patients managed expectantly and by the study by Pascual et al being an extreme low-risk population. [17,52] The classification performance of the Simple Rules (inconclusive cases classified as malignant) in our study (sensitivity 0.90, specificity 0.85) is reasonably similar to that reported in published meta-analyses (sensitivity and specificity 0.93 and 0.81,[50] sensitivity and specificity 0.94 and 0.76, [51] sensitivity and specificity 0.93 and 0.80 [9]) and in the international multicenter study by van Calster et al (sensitivity and specificity 0.90 and 0.87 [22]).

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, specificity 0.89).[9] The classification performance of subjective assessment is heavily dependent on the experience of the ultrasound examiner.[53] 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 experience, 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.

*Implications in clinical practice*

The good performance of the IOTA models in our study, which includes also ultrasound examiners with little ultrasound experience and both 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 two-step strategy. The two-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 two-step strategy is easier to use than ADNEX while still offering the advantage of providing an estimate of the likelihood of four types of malignancy.

*Future perspectives*

Prospective studies including a very large number of ultrasound examiners with limited experience are needed to confirm our results. It would be important to investigate the effect of using the IOTA models in impact studies. [54] Such studies will show whether use of IOTA models in daily practice improve decision making and, ultimately, patient outcome.

*Conclusions*

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

**References**

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12-49. doi: 10.3322/caac.21820.

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(2):286-90. doi: 10.1016/j.ygyno.2012.04.012.

3. Engelen MJ, Kos HE, Willemse PH, et al. Surgery by consultant gynecologic oncologists improves survival in patients with ovarian carcinoma. *Cancer*. 2006;106(3):589-98. doi: 10.1002/cncr.21616.

4. Vernooij F, Heintz AP, 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. *J Natl Cancer Inst*. 2008;100(6):399-406. doi: 10.1093/jnci/djn033.

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(3):448-458. doi: 10.1016/S1470-2045(18)30837-4.

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(1):148-168. doi: 10.1002/uog.23635.

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(4):357-65. doi: 10.1046/j.0960-7692.2001.00500.x.

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(1):91-104. doi: 10.1016/j.bpobgyn.2003.09.009.

9. Meys EM, Kaijser J, Kruitwagen RF, et al. Subjective assessment versus ultrasound models to diagnose ovarian cancer: A systematic review and meta-analysis. *Eur J Cancer*. 2016 May;58:17-29. doi: 10.1016/j.ejca.2016.01.007.

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. *Br J Obstet Gynaecol*. 1990;97(10):922-9. doi: 10.1111/j.1471-0528.1990.tb02448.x.

11. https://kunskapsbanken.cancercentrum.se/diagnoser/aggstockscancer-epitelial/vardprogram/diagnostik/

12. https://www.legeforeningen.no/contentassets/75de8a48892f4cdc8db45fc7804d369c/benigne-ovarialcyster_2024.pdf

13. https://static1.squarespace.com/static/5467abcce4b056d72594db79/t/58ed3a3ddb29d654c6cd634c/1491941952677/Cysteguideline.pdf.

14. Royal College of Obstetricians and Gynaecologists. The Management of Ovarian Cysts in Postmenopausal Women. Green-top Guideline No. 34. July 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 Oct;25(10):1371-1386. doi: 10.1016/S1470-2045(24)00406-6.

16. Ameye L, Timmerman D, Valentin L, et al. Clinically oriented three-step strategy for assessment of adnexal pathology. *Ultrasound Obstet Gynecol*. 2012;40(5):582-91. doi: 10.1002/uog.11177.

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 Feb;61(2):231-242. doi: 10.1002/uog.26080.

18. Timmerman D, Testa AC, Bourne T, et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol*. 2008;31(6):681-90. doi: 10.1002/uog.5365.

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(34):8794-801. doi: 10.1200/JCO.2005.01.7632.

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(4):424-437. doi: 10.1016/j.ajog.2016.01.007.

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. doi: 10.1136/bmj.g5920.

22. 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. doi: 10.1136/bmj.m2614.

23. 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(5):668-676. doi: 10.1002/uog.24777.

24. 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(6):2010. doi: 10.3390/jcm9062010.

25. 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: Simple Rules and Assessment of Different NEoplasias in the adneXa Model. *J Ultrasound Med*. 2022;41(2):471-482. doi: 10.1002/jum.15728.

26. 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. *Womens Health Rep* (New Rochelle). 2023;4(1):202-210. doi: 10.1089/whr.2022.0072.

27. 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* (Basel). 2022;11(1):8. doi: 10.3390/healthcare11010008.

28. 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. doi: 10.3389/fonc.2021.673722.

29. 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(6):778-783. doi: 10.1002/uog.15963.

30. 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(2):226-34. doi: 10.1002/uog.7636.

31. 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. doi: 10.1136/bmj.c6839.

32. 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(5):542-8. doi: 10.1038/bjc.2016.227.

33. 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(1-2):132-138. doi: 10.1159/000513517.

34. 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;9(1):E11-E17. doi: 10.1055/a-2044-2855.

35. 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(3):304-318. doi: 10.4103/ijri.IJRI_279_20.

36. https://iotaplus.org/en/certified-members.

37. European Federation of Societies for Ultrasound in Medicine and Biology. *Ultraschall Med* 2006;27(1): 79-95 doi: 10.1055/s-2006-933605

38. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I; International Ovarian Tumor Analysis (IOTA) Group. 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(5):500-5.

39. Prat J; FIGO Committee on Gynecologic Oncology. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. *Int J Gynaecol Obstet*. 2014;124(1):1-5. doi: 10.1016/j.ijgo.2013.10.001.

40. 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(2):e28. doi: 10.2196/medinform.3251.

41. 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(19):4230-4251. doi: 10.1002/sim.9025.

42. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ*. 2017;356:i6460. doi: 10.1136/bmj.i6460.

43. 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. doi: 10.1016/j.jclinepi.2015.12.005.

44. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565-74. doi: 10.1177/0272989X06295361.

45. Lin L, Chu H. Meta-analysis of Proportions Using Generalized Linear Mixed Models. *Epidemiology*. 2020;31(5):713-717. doi: 10.1097/EDE.0000000000001232.

46. Van Calster B, Vergouwe Y, Looman CW, 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(10):761-70. doi: 10.1007/s10654-012-9733-3.

47. 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(23):2610-26. doi: 10.1002/sim.5321.

48. 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(1):e000817. doi: 10.1136/bmjmed-2023-000817.

49. 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(3):397-402. doi: 10.1002/uog.15854.

50. 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(3):449-62. doi: 10.1093/humupd/dmt059.

51. 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.* 2018;22(44):1-264. doi: 10.3310/hta22440.

52. 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(3):395-404. doi: 10.1002/uog.27642.

53. 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(1):11-6. doi: 10.1046/j.1469-0705.1999.13010011.x.

54. Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ*. 2009;338:b606. doi: 10.1136/bmj.b606.

## Tables

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

| | *Median (IQR), or n (%), Range* |
|---|---|
| Patient age at recruitment (years) | 52 (IQR 40-62)<br>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)<br>Range: 9-400 |
| Largest diameter of largest solid component (mm)* | 40 (IQR 16-68)<br>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)<br>Range: 1-12000 |
| Colour score of intratumoural 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† | |
| 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
*For tumors with a solid component
† specific histological diagnoses are shown in Supplementary Table S2.

**Table 2.** Calibration in terms of observed over expected ratio based on meta-analysis of data from

| **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 is underestimating the average risk. A value <1 means that the model is over-estimating the average risk.
CI; Confidence Interval, SRRisk; Simple Rules Risk Model, ADNEX; Assessment of Different NEoplasias in the adnexa.

**Table 3.** Polytomous discrimination index (PDI) of Assessment of Different NEoplasias in the adneXa (ADNEX), and of the two-step strategy (pooled analysis).

| | **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) |

**Table 4**. Tumor outcome and percentage of missing CA 125 values for all pre-specified subgroups

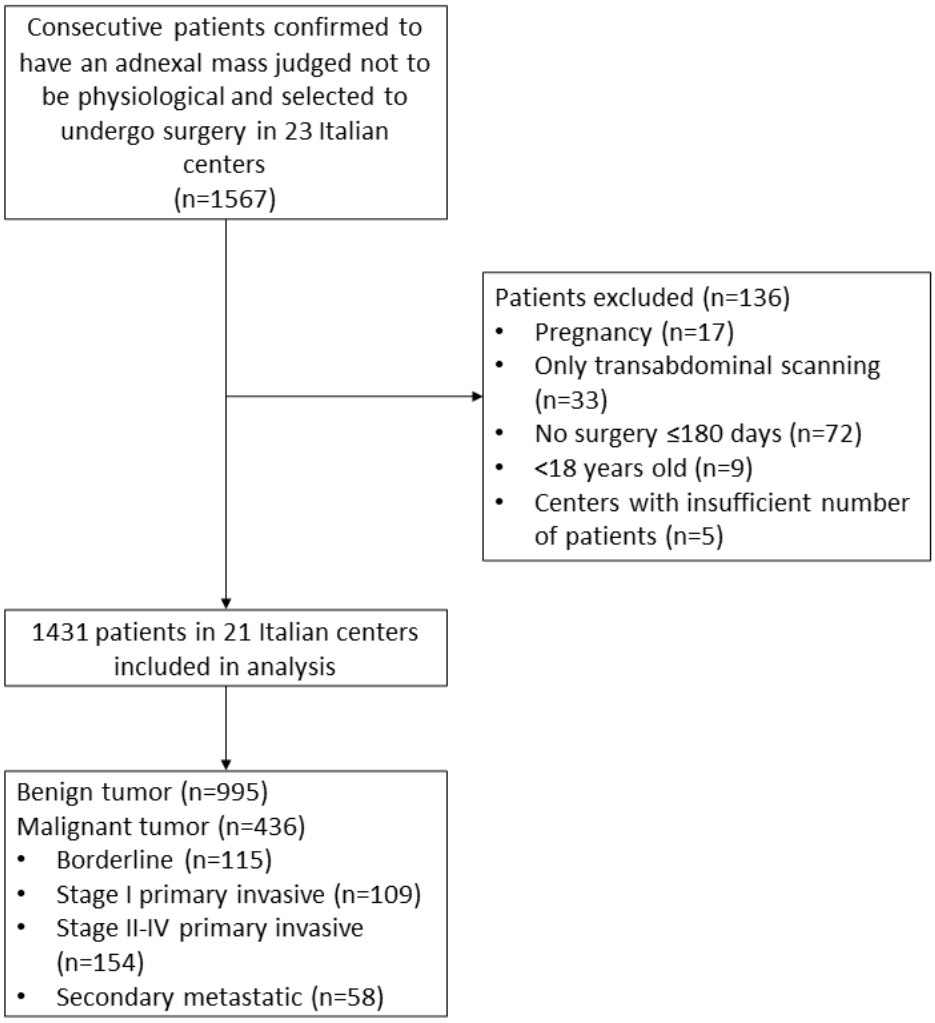
| | | Outcome | | |
|---|---|---|---|---|
| **Subgroup** | **N** | **Benign** | **Malignant** | **Missing CA125** |
| Postmenopausal | 745 | 444 (60) | 301 (40) | 176 (24) |
| Premenopausal | 686 | 551 (80) | 135 (20) | 218 (32) |
| Oncology center | 817 | 583 (64) | 331 (36) | 219 (27) |
| Other center | 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) |
| Level2 | 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

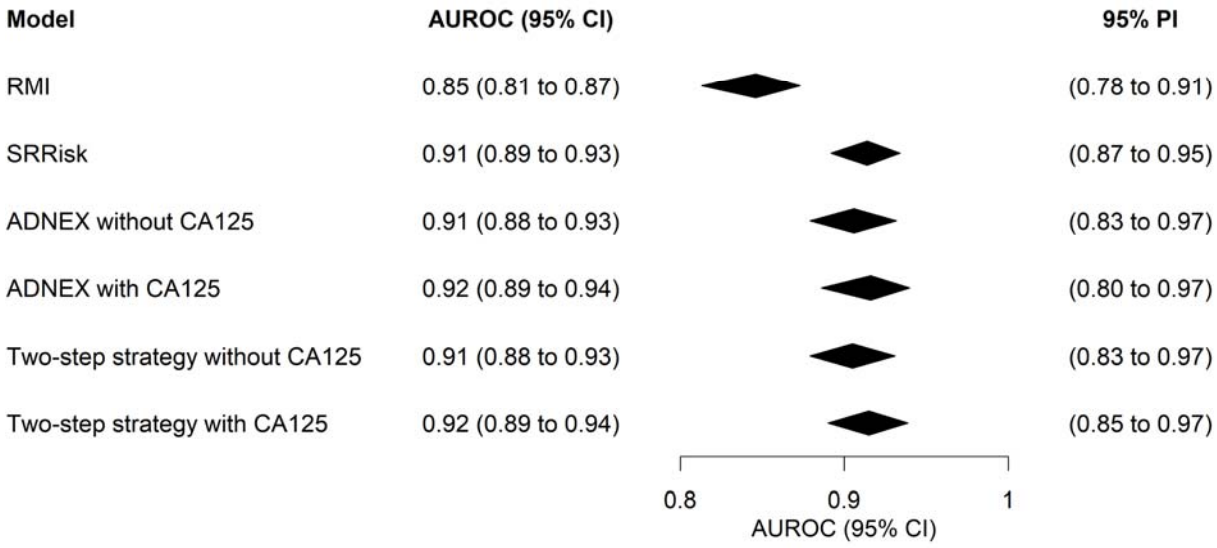
852
853
854
855

856 **Figure 1.** Flowchart of the patients included for the analysis.



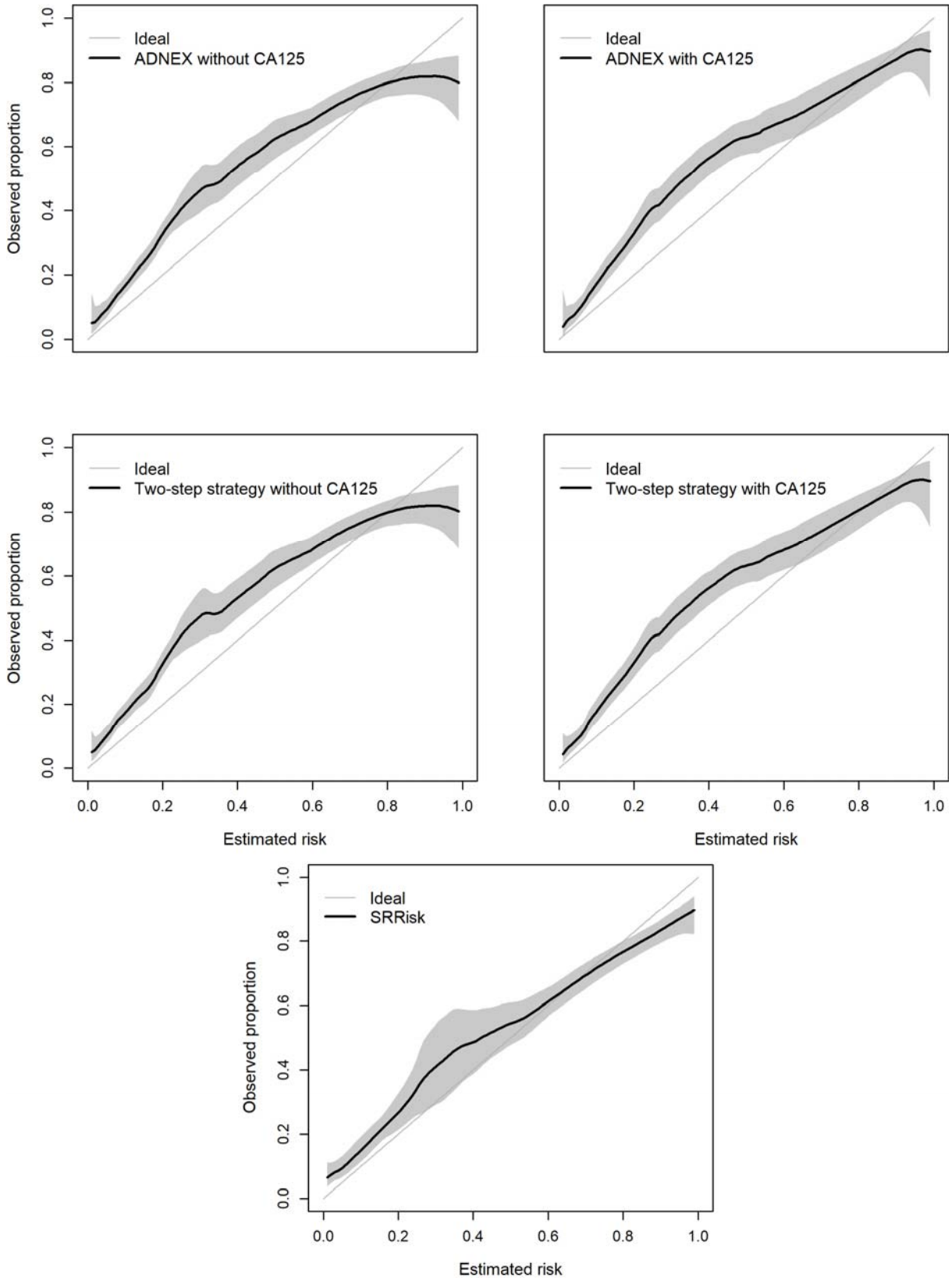

857
858
859

860 **Figure 2.** Summary forest plot of area under the receiver operating characteristic curve (AUROC)
861 based on meta-analysis of data from 21 centers. CI; Confidence Interval, PI; Prediction Interval.



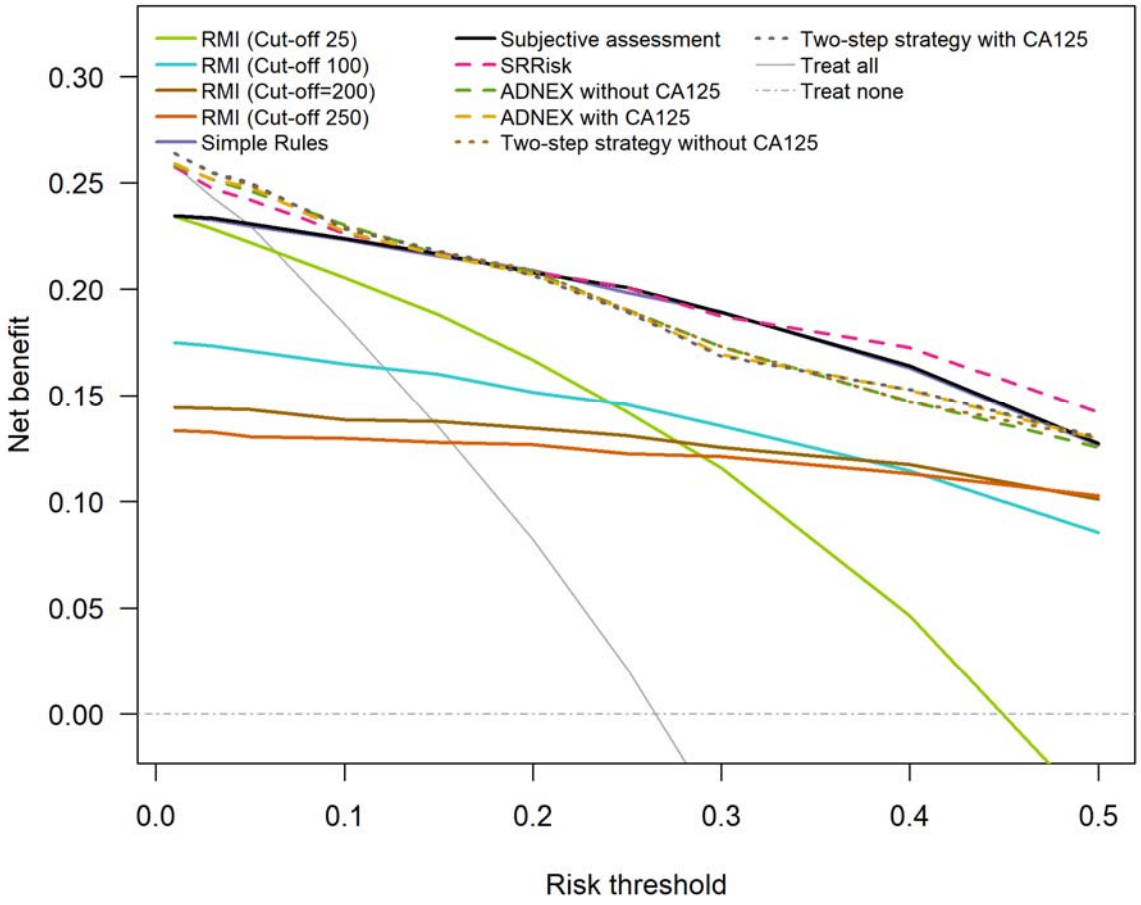

862
863
864

**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 four groups: Santorso, Foggia, Treviso (group 1); Messina, Carpi, Montebelluna (group 2); Verona, Firenze, Padova, Roma (group 3); Bari B, Asti, Bolzano (group 4).The curves were obtained with meta-analysis of center-specific curves from eight centers and from the four groups. SRRisk; Simple Rules Risk Model, ADNEX; Assessment of Different NEoplasias in the adnexa.



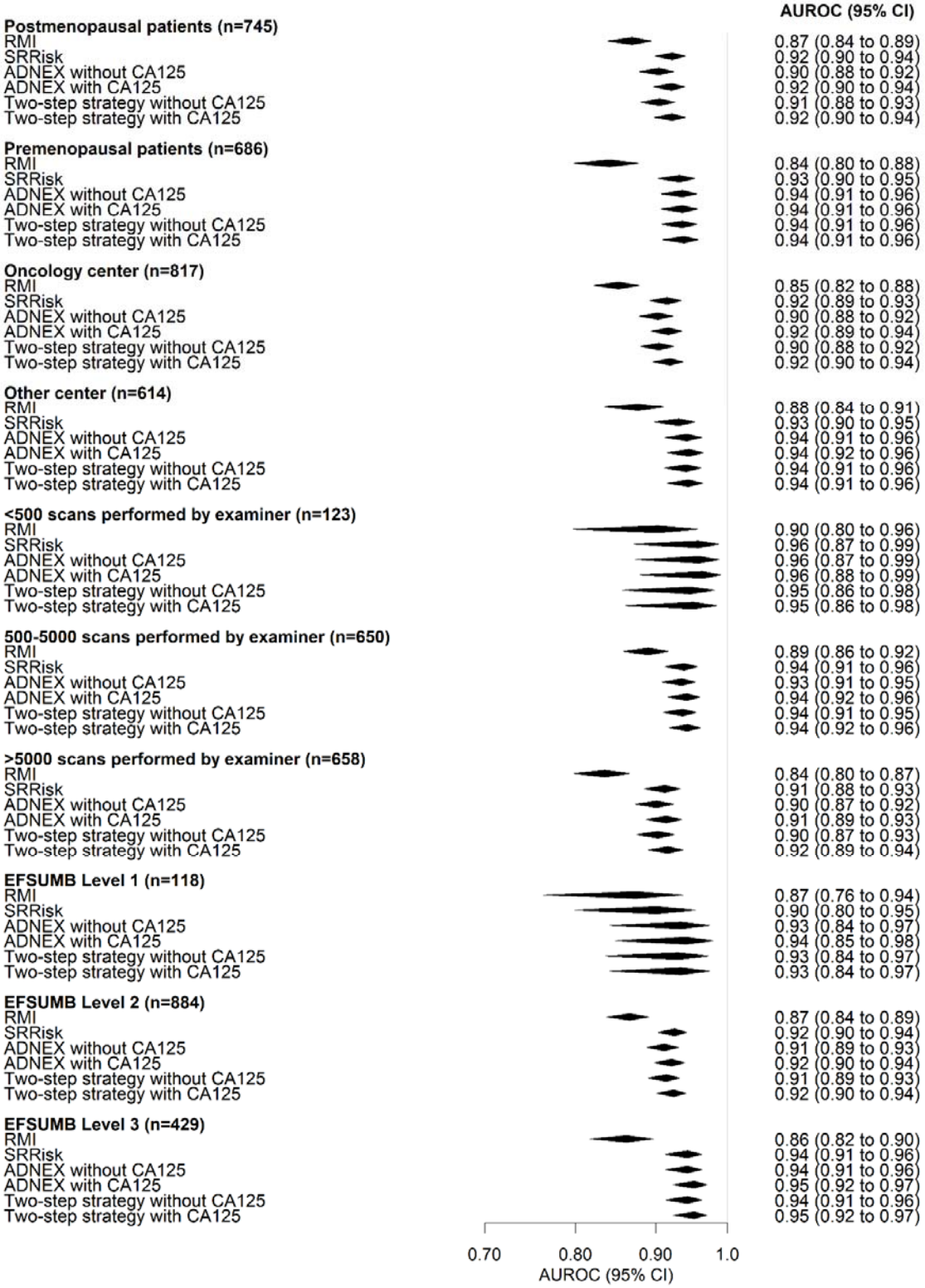

871
872
873

874 **Figure 4.** Decision curves for risk models, Risk of Malignancy Index (RMI), Simple Rules and
875 subjective assessment based on meta-analysis of data from 21 centers. The curves show net benefit
876 at several thresholds between 1% and 50%. A model is clinically useful if it is superior to both treat
877 all and treat none. At risk thresholds below 15%, using RMI at cutoff 200 is worse than treating
878 everyone (i.e.,worse than referring all women with an adnexal mass to a gynecological oncology
879 center).




880
881
882

883 **Figure 5**. Forest plot of area under the receiver operating characteristic curve (AUROC) for
884 prespecified subgroups (pooled data). CI; Confidence interval, PI; Predicition interval, EFSUMB;
885 European Federation of Societies for Ultrasound in Medicine and Biology.
886



887