Landolfo Chiara (Orcid ID: 0000-0001-9808-7957)
Froyman Wouter (Orcid ID: 0000-0002-1398-9124)
Sladkevicius Povilas (Orcid ID: 0000-0001-9563-8135)
Kudla Marek Jerzy (Orcid ID: 0000-0003-0928-3433)
Alcazar Juan Luis (Orcid ID: 0000-0002-9700-0853)
Guerriero Stefano (Orcid ID: 0000-0002-1359-7155)
Fischerova Daniela (Orcid ID: 0000-0002-7224-3218)
Valentin Lil (Orcid ID: 0000-0002-3830-6414)

Benign descriptors and ADNEX in two-step strategy to estimate risk of malignancy in ovarian tumors: retrospective validation on IOTA 5 multicenter cohort

C. Landolfo[1,2], T. Bourne[1,3,4], W. Froyman[1,3], B. Van Calster[1,5], J. Ceusters[1,6], A. C. Testa[2,7], L. Wynants[1,8], P. Sladkevicius[9,10], C. Van Holsbeke[11], E. Domali[12], R. Fruscio[13], E. Epstein[14,15], D. Franchi[16], M. J. Kudla[17], V. Chiappa[18], J. L. Alcazar[19], F. P. G. Leone[20], F. Buonomo[21], M. E. Coccia[22], S. Guerriero[23], N. Deo[24], L. Jokubkiene[9,10], L. Savelli[25], D. Fischerova[26], A. Czekierdowski[27], J. Kaijser[28], A. Coosemans[6], G. Scambia[2,7], I. Vergote[3,6], D. Timmerman[1,3]* and L. Valentin[9,10]*

[1]Department of Development and Regeneration, KU Leuven, Leuven, Belgium

[2]Department of Woman, Child and Public Health, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, Italy

[3]Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium

[4]Queen Charlotte's and Chelsea Hospital, Imperial College, London, UK

[5]Department of Biomedical Data Sciences, Leiden University Medical Centre (LUMC), Leiden, Netherlands

[6]Laboratory of Tumor Immunology and Immunotherapy, Department of Oncology, Leuven Cancer Institute, KU Leuven, Leuven, Belgium

[7]Dipartimento Universitario Scienze della Vita e Sanità Pubblica, Università Cattolica del Sacro Cuore, Rome, Italy

[8]Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands

[9]Department of Obstetrics and Gynecology, Skåne University Hospital, Malmö, Sweden

[10]Department of Clinical Sciences Malmö, Lund University, Sweden

[11]Department of Obstetrics and Gynecology, Ziekenhuis Oost-Limburg, Genk, Belgium

[12]First Department of Obstetrics and Gynecology, Alexandra Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece

[13]Clinic of Obstetrics and Gynecology, University of Milan-Bicocca, San Gerardo Hospital, Monza, Italy

[14]Department of Clinical Science and Education, Karolinska Institutet, Stockholm, Sweden

[15]Department of Obstetrics and Gynecology, Södersjukhuset, Stockholm, Sweden

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.26080

[16]Preventive Gynecology Unit, Division of Gynecology, European Institute of Oncology IRCCS, Milan, Italy
[17]Department of Perinatology and Oncological Gynecology, Faculty of Medical Sciences, Medical University of Silesia, Katowice, Poland
[18]Department of Gynecologic Oncology, National Cancer Institute of Milan, Milan, Italy
[19]Department of Obstetrics and Gynecology, Clinica Universidad de Navarra, School of Medicine, Pamplona, Spain
[20]Department of Obstetrics and Gynecology, Biomedical and Clinical Sciences Institute L. Sacco, University of Milan, Milan, Italy
[21]Institute for Maternal and Child Health, IRCCS 'Burlo Garofolo', Trieste, Italy
[22]Department of Obstetrics and Gynecology, University of Florence, Florence, Italy
[23]Department of Obstetrics and Gynecology, University of Cagliari, Policlinico Universitario Duilio Casula, Cagliari, Italy
[24]Department of Obstetrics and Gynecology, Whipps Cross Hospital, London, UK
[25]Gynecology and Physiopathology of Human Reproduction Unit, S. Orsola-Malpighi Hospital of Bologna, Bologna, Italy
[26]Gynecologic Oncology Centre, Department of Obstetrics and Gynecology, First Faculty of Medicine, Charles University and General University Hospital in Prague
[27]First Department of Gynecological Oncology and Gynecology, Medical University of Lublin, Lublin, Poland
[28]Department of Obstetrics and Gynecology, Ikazia Hospital, Rotterdam, Netherlands

*Contributed equally

Corresponding author: Prof. L. Valentin

Department of Obstetrics and Gynecology, Skåne University Hospital, Malmö, Sweden

E-mail: lil.valentin@med.lu.se

Short title: IOTA two-step strategy

Keywords: ultrasonography; ovarian neoplasms; validation study; IOTA; ADNEX model; benign simple descriptors

ORCID IDs

Chiara Landolfo: 0000-0001-9808-7957

Valentina Chiappa: 0000-0002-1811-209X

Laure Wynants: 0000-0002-3037-122X

Jeroen Kaijser: 0000-0003-2047-1988

Francesca Buonomo: 0000-0002-6587-2622

Francesco Leone: 0000-0003-3839-6779

Ligita Jokubkiene: 0000-0002-0758-3507

An Coosemans: 0000-0002-7321-4339

Elizabeth Epstein: 0000-0003-2298-7785

Artur Czekierdowski: 0000-0001-6481-4271

Luca Savelli: 0000-0003-0961-7296

Daniela Fischerova: 0000-0002-7224-3218

Vergote Ignace: 0000-0002-7589-8981

Ben Van Calster: 0000-0003-1613-7450

Antonia Testa: 0000-0003-2217-8726

Dirk Timmerman: 0000-0002-3707-6645

Tom Bourne: 0000-0003-1421-6059

Maria Elisabetta Coccia: 0000-0001-5294-129

Povilas Sladkevicious: 0000-0001-9563-8135

Ekaterini Domali: 0000-0001-8899-3040

Wouter Froyman: 0000-0002-1398-9124

Juan Louis Alcazar: 0000-0002-9700-0853

Marek Kudla: 0000-0003-0928-3433

Caroline Van Holsbeke: 0000-0002-9128-1367

Franchi Dorella: 0000-0002-3950-5538

Lil Valentin: 0000-0002-3830-6414

Stefano Guerriero: 0000-0002-1359-7155

CONTRIBUTION

What are the novel findings of this work?

This study is the first to validate the modified IOTA Benign Simple Descriptors (BD) and a two-step strategy to estimate the risk of malignancy in adnexal masses. BDs are used as the first step, followed by ADNEX if BDs do not apply. The strategy had excellent discriminative ability but slightly underestimated the risk of malignancy.

What are the clinical implications of this work?

The two-step strategy is suitable for clinical use. A large proportion of adnexal masses can be classified as benign by the BDs without computer support. For the remaining masses, malignancy risk can be calculated using ADNEX. An ADNEX calculator is available online, as an application for smartphones and is embedded in many ultrasound machines.

ABSTRACT

Objective: Previous work suggested that the ultrasound-based benign Simple Descriptors can reliably exclude malignancy in a large proportion of women presenting with an adnexal mass. We aim to validate a modified version of the Benign Simple Descriptors (BD), and we introduce a two-step strategy to estimate the risk of malignancy: if the BDs do not apply, the ADNEX model is used to estimate the risk of malignancy.

Methods: This is a retrospective analysis using the data from the 2-year interim analysis of the IOTA5 study, in which consecutive patients with at least one adnexal mass were recruited irrespective of subsequent management (conservative or surgery). The main outcome was classification of tumors as benign or malignant, based on histology or on clinical and ultrasound information during one year of follow-up. Multiple imputation was used when outcome based on follow-up was uncertain according to predefined criteria.

Results: 8519 patients were recruited at 36 centers between 2012 and 2015. We included all masses that were not already in follow-up at recruitment from 17 centers with good quality surgical and follow-up data, leaving 4905 patients for statistical analysis. 3441 (70%) tumors were benign, 978 (20%) malignant, and 486 (10%) uncertain. The BDs were applicable in 1798/4905 (37%) tumors, and 1786 (99.3%) of these were benign. The two-step strategy based on ADNEX without CA125 had an area under the receiver operating characteristic curve (AUC) of 0.94 (95% CI, 0.91-0.95). The risk of malignancy was slightly underestimated, but calibration varied between centers. A sensitivity analysis in which we expanded the definition of uncertain outcome resulted in 1419 (29%) tumors with uncertain outcome and an AUC of the two-step strategy without CA125 of 0.93 (95% CI, 0.91-0.95).

Conclusion: A large proportion of adnexal masses can be classified as benign by the BDs. For the remaining masses the ADNEX model can be used to estimate the risk of malignancy. This two-step strategy is convenient for clinical use.

INTRODUCTION

Ovarian cancer is the fifth leading cause of cancer death among women in developed countries. Patients with ovarian cancer treated in tertiary oncology referral centers have a better prognosis than those managed in general gynecology departments[1–4]. Correct diagnosis is important to be able to offer optimal treatment.

To help clinicians decide on appropriate management, mathematical models to predict malignancy in adnexal masses have been developed on cohorts of patients that underwent surgery. A well-known model is the Risk of Malignancy Index (RMI)[5]. The International Ovarian Tumor Analysis (IOTA) group created and validated four models to estimate the risk of malignancy in adnexal masses: logistic regression model 1 (LR1), logistic regression model 2 (LR2), Simple Rules risk model (SRRisks), and Assessment of Different NEoplasias in the adneXa (ADNEX)[6–10]. Systematic reviews and prospective cohort studies have shown that IOTA models discriminate better between benign and malignant tumors than all other models including the RMI[6,11–14]. The ADNEX model uses simple predictor variables and calculates the risk of four types of malignancy[7].

Some adnexal lesions can be easily classified as benign or malignant using the IOTA "Simple Descriptors". These are based on easily recognizable ultrasound features and do not require access to a computer[15]. If a benign Simple Descriptor applies to a tumor selected for surgery, the tumor is almost certainly benign (>99%), while >92% of tumors to which a malignant Simple Descriptor applies are malignant[6,15]. In clinical practice, it would be logical to first apply the benign Simple Descriptors. If one of these applies, the mass could be classified as benign (risk of malignancy <1%), while if none applies, a mathematical model could be used to estimate the risk of malignancy. To the best of our knowledge, such a two-step strategy has not been suggested before, nor has it been validated either in masses removed by surgery or in masses managed conservatively.

The primary aim of this study is to validate the diagnostic performance of the benign Simple Descriptors and of a two-step strategy, i.e. benign Simple Descriptors followed by ADNEX if benign Simple Descriptors do not apply, when used in both surgically and conservatively managed adnexal masses.

METHODS

Study design

This was a retrospective analysis of the interim data from the IOTA phase 5 study (IOTA5), an international multicenter prospective cohort study[16,17]. Consecutive patients with at least one adnexal tumor examined with transvaginal ultrasonography were included. Surgery or conservative management was suggested by the ultrasound examiner based on the ultrasound appearance of the tumor (pattern recognition), symptoms, and evolution of the tumor over time. In the IOTA5 study, patients were recruited until March 2015. However, patient follow-up will continue until each conservatively managed patient has been followed-up for at least five years. The interim analysis includes patients enrolled between January 1st, 2012 and March 1st, 2015, and follow-up data until June 30th, 2017. A total of 36 centers in 14 countries participated in the study, both oncology referral centers (tertiary centers with a specific gynecologic oncology unit) and other types of centers. Approvals from the ethics committee of the University Hospitals Leuven as the coordinating center (B32220095331/S51375) and from the local ethics committee of each contributing center were obtained (ethical approval numbers are listed in Table S1). The IOTA5 study protocol can be found at ClinicalTrials.gov (NCT01698632). The current report is written in accordance with Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines[18].

Inclusion criteria

Patients were eligible for inclusion if they were at least 18 years old and had at least one adnexal (ovarian/para-ovarian or tubal/para-tubal) tumor at ultrasonography. We used the IOTA definition of adnexal tumor (lesion), i.e. “the part of an ovary or an adnexal mass that is judged from assessment of ultrasound images to be inconsistent with normal physiological function”[19].

Exclusion criteria

Cysts judged to be physiological (follicular cysts, corpus luteum cysts) with a largest diameter <3 cm were not eligible for inclusion in IOTA 5. Denial or withdrawal of informed consent were other exclusion criteria. Pregnancy was not an exclusion criterion. For the analysis of this study, patients with adnexal tumors already diagnosed and in follow-up in the participating center before enrolment in the IOTA 5 study were excluded.

Ultrasound examination and CA125 measurement

At inclusion, ultrasound examiners performed a standardized transvaginal ultrasound examination and registered clinical information following a research protocol. By design the ultrasound examiners were blinded to the outcome. They were not actively blinded to clinical information, nor to results of

biomarkers or other imaging, e.g. computer tomography, that might have been performed before the ultrasound examination. All ultrasound examiners (n = 77) had passed the IOTA certification test (iotagroup.org/certified-members). Most scans were performed by level II or III examiners, very few were performed by level I examiners (level defined by the European Federation of Societies of Ultrasound in Medicine and Biology, EFSUMB20). IOTA terminology was used to describe the ultrasound findings[19]. Information on predefined ultrasound variables was collected for each patient (Table S2). Using subjective assessment (pattern recognition), ultrasound examiners classified each tumor as benign, borderline or malignant and specified the degree of certainty with which the diagnosis was made (certain, probable, uncertain). The ultrasound diagnoses were based on knowledge of the typical ultrasound appearance of benign, borderline and malignant lesions and that of different types of specific adnexal pathology[21]. If there were multiple masses, the one with the most complex ultrasound morphology was registered by the ultrasound examiner as the dominant tumor. The dominant tumor was used in our statistical analysis. At follow-up visits, ultrasound examination was performed following the same protocol as at the inclusion scan, and clinical information, including information on symptoms, was obtained. At each examination, the ultrasound examiner proposed management (surgical removal or follow-up) based on the ultrasound diagnosis and the patient´s symptoms. However, the final decisions about management were made by the referring clinicians taking into account clinical symptoms, ultrasound findings and findings of other imaging modalities such as computer tomography or magnetic resonance imaging, tumor markers, and patients' preferences.

Conservative management comprised clinical and ultrasound follow-up at intervals of three months, six months, and 12 months, and then every 12 months thereafter.

Measurement of serum CA125 was encouraged, but it was not an inclusion criterion for the study. Measurements of CA125 were performed according to local practice in each center.

Data collection and cleaning

Patient data were registered on a secure electronic platform (IOTA5 Study Screen; astraia Software, Munich, Germany). A unique identifier code was automatically assigned to each patient. All data communications were encrypted to guarantee data security. Data cleaning was performed by a team of ultrasound examiners and biostatisticians. It included queries to local investigators to amend inconsistencies and complete missing data. A standardized questionnaire (Appendix S1) for patients and/or managing clinicians was used at the local centers to retrieve missing information. Before analyzing our data, we defined the criteria for a study center to be included in our analysis. For a center to be included, we required it to have recruited at least 50 patients, to have recruited patients consecutively irrespective of suggested management (surgery or conservative management with follow-up), and to have good quality follow-up data for at least 70% of the recruited patients. We defined good follow-up data as a recorded study outcome (surgery at any point, spontaneous resolution of the mass,

or patient death) or a last follow-up visit 10 months or later after inclusion. The 70% cutoff was chosen arbitrarily, because it seemed reasonable to members of the IOTA Steering Committee (details in Appendix S2).

The modified benign Simple Descriptors and the two-step strategy

We modified *à priori* the original benign Simple Descriptors[15] by requiring the largest diameter of the tumor to be <10 cm for all four benign descriptors instead of only for the third benign descriptor (Figure 1). We refer to these descriptors as modified Benign Simple Descriptors (Benign Descriptors, BDs). The size criterion was added to decrease the likelihood of a malignant tumor being misclassified as benign. Based on data from the IOTA phase 1-3 studies (n=5914)[6], 1618 tumors (27%) fulfilled the criteria of a benign Simple Descriptor, and 11 of these (0.7%) were malignant. Among the same 5914 tumors, 1427 (24%) tumors fulfilled the criteria of a modified benign Simple Descriptor (BD), and six (0.4%) of these were malignant (unpublished data).

The malignant Simple Descriptors[15] are not used in the two-step strategy. As a first step in the two-step strategy the BDs are applied. When the BDs do not apply, the second step is to apply ADNEX. ADNEX calculates the probability of five outcome categories: benign, borderline, stage I primary invasive ovarian malignancy, stage II-IV primary invasive ovarian malignancy, and a metastasis in the ovary from another primary origin (e.g. breast cancer or colon cancer)[7]. ADNEX uses three clinical and six ultrasound predictors: type of center (oncology center vs other), patient age, CA125 level, maximum diameter of the lesion, proportion of solid tissue, number of papillary projections, presence of >10 cyst locules, presence of acoustic shadows, and presence of ascites. ADNEX can also be used without CA125. Details on the ADNEX model are provided in Appendix S3. Model predictions are based on information obtained at the inclusion scan and hence are blinded to the outcome.

Reference standard

The reference standard refers to the nature of the adnexal tumor (benign or malignant) at inclusion. Borderline tumors were classified as malignant. Each adnexal mass was classified as benign or malignant based on histology, if the tumor was surgically removed, otherwise on the results of follow-up examinations (see below). The histology of the surgically removed tumor was determined at the local center. Central pathology review was not performed, because we found little differences between local and central pathology reports in a previous IOTA study[8]. 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 classification of malignant tumors was done according to the International Federation of Gynecology and Obstetrics[22]. If the tumor was not surgically removed, it was classified as benign or malignant based on clinical and ultrasound findings during 12 months +/- two months of follow-up (i.e., minimal follow-up time to assign an outcome was 10 months). Different scenarios were possible: some patients underwent surgery without

follow-up, others were managed conservatively with or without surgery later. For some patients we have no information after the inclusion visit. If data to classify the tumor as benign or malignant at inclusion was not available, the outcome was classified as uncertain. Table 1 describes the criteria for classifying tumors as benign, malignant or uncertain.

Study endpoints

In line with the study objectives, the main study endpoints are 1) the percentage of tumors to which the BDs apply, 2) the percentage of malignant tumors among lesions to which the BDs apply, and 3) the diagnostic performance in terms of discrimination and calibration of the two-step strategy. Secondary study endpoint is the discriminative ability of ADNEX (with and without CA125) when applied only on tumors to which the BDs do not apply.

Statistical analysis

A summary of the statistical analysis is provided below. Details on statistical analysis and discussion of sample size can be found in Appendix S4-6. The statistical analysis was performed with R version 3.5.1.

We calculated the percentage of patients to which the BDs applied, and the prevalence of malignancy in tumors to which the BDs applied.

To assess performance of the two-step strategy, we needed risk estimates for each of the five tumor outcomes when the BDs applied. Appendix S5 describes how this was done. We evaluated discrimination between benign and malignant tumors using the area under the receiver operating characteristic curve (AUC). To evaluate calibration of the estimated risk of malignancy, we calculated the calibration intercept and slope using a logistic recalibration model23. Clinical utility by using decision curve analysis was assessed by calculating Net Benefit at thresholds for estimated risks of malignancy between 5% and 50% to decide which patients to refer to specialized oncological care24.

For the two-step strategy, we further assessed the AUC for each pair of tumor subtype, the Polytomous Discrimination Index (PDI) as a multiclass AUC, and calibration for the estimated risks of each of the five tumor subtypes[25,26].

For the percentage of patients to which the BDs applied, AUC for benign vs malignant tumors, calibration of the risk of malignancy, and decision curve analysis, we addressed heterogeneity between centers. This was done by calculating center-specific performance and combining the results using meta-analysis (see Appendix S4)[23,24,27]. Heterogeneity in the AUC was quantified using 95% prediction intervals (PI)[28], which describe the range of AUC values that can be expected in a new center. Because the number of malignant tumors was too low, meta-analysis was not possible for the prevalence of malignancy in tumors to which the BDs applied, AUC for each pair of tumor subtype, PDI, and

calibration for each tumor subtype. For these analyses, data from all centers were pooled. For the percentage of patients to which the BDs applied, we performed both a meta-analysis and a pooled analysis.

Subgroup analyses were performed for menopausal status and type of center.

Methods to address potential sources of bias

We implemented several procedures to reduce potential bias. First, we followed a prespecified statistical analysis plan to avoid selecting analyses based on results. Second, to handle differential verification, we used prespecified criteria to determine whether tumor outcome was benign, malignant, or uncertain (Table 1). Third, the primary analysis includes all patients after multiple imputation of missing CA125 levels and uncertain outcomes (Appendix S4). Excluding participants with uncertain outcome leads to partial verification bias, excluding participants with missing CA125 leads to selection bias[18,27,29,30]. Multiple imputation is a recommended approach to avoid such exclusions[30]. Fourth, we performed a prespecified sensitivity analysis in which we expanded the definition of uncertain outcomes to include all groups in which subjective assessment of ultrasound images was used to label outcomes as benign or malignant (B2, M2-3, and U1-4 in Table 1). This was done to address possible optimistic bias due to differential verification. Fifth, we used prespecified criteria for data quality in order to include only data from centers with consecutive inclusion and sufficiently complete and accurate data (Appendix S2). This may limit potential attrition bias by avoiding exclusions on patient level (instead, we excluded entire centers) and limits the number of uncertain outcomes. Finally, an additional prespecified analysis was performed in which masses with uncertain outcome as per Table 1 were excluded. This was done for completeness only, because exclusions based on missing data result in high risk of bias[18,30].

RESULTS

Patient flow is shown in Figure S1. A total of 8519 patients recruited at 36 centers were included in the interim dataset of IOTA5 (Table S3). Twenty-five patients were excluded due to withdrawal of consent. Another 2777 patients from 19 centers were excluded from the primary analysis: one center stopped participation, seven centers recruited <50 participants, three centers were excluded due to non-consecutive recruitment and eight centers due to suboptimal data quality. Suboptimal data quality is explained by lack of staff (three centers), problems with information technology (two centers) or difficulties with making patients return for planned follow-up visits (four centers). Of the 5717 patients in the remaining 17 centers, 812 (14%) patients had a mass that was already being followed in the recruitment center before inclusion. Therefore, 4905 patients were included in our primary analysis (Table S1). In 4151 (85%) of the 4905 women the ultrasound examiner´s suggestion for management was followed, in 445 (9%) it was not followed, and in 309 (6%) the actual management was unknown.

Patient and tumor characteristics are shown in Table 2. Median age of the 4905 patients was 48 years (interquartile range 36-62, range 18-98), 2151 patients (44%) were postmenopausal, 2140 (44%) had a dominant mass that was a unilocular cyst, and 1734 (35%) had a dominant mass containing solid components. Median maximum lesion diameter was 55 mm (interquartile range 38-83, range 7-751), and 2031 masses (41%) had no detectable blood flow on color or power Doppler (color score 1). Information on CA125 was missing in 2620 of the 4905 (53%) patients. Missing CA125 values were less common for patients who underwent surgery (32%) or who had a tumor considered probably malignant (23%) or certainly malignant (14%) (Table S4). In all, 3441 (70%) tumors were benign, 978 (20%) were malignant (borderline or invasive), and for 486 (10%) tumors the outcome was uncertain (Table S1). Uncertain outcome was explained by loss to follow-up (n=432), or by conflicting information during follow-up (n=54) (Table 1). Loss to follow-up was less common when surgery was suggested (5%) than when conservative management was suggested (13%) and less common when the diagnosis based on subjective assessment was benign (10%) than when it was uncertain or malignant (5%) (Table S5). A lower proportion of tumors in this study than in the development dataset of ADNEX manifested malignant ultrasound features. This is because the development set included only patients that underwent surgery (see details elsewhere[17]).

The BDs were applicable in 37% (1798/4905) of tumors (pooled analysis). Center-specific results and the result of meta-analysis are shown in Table S6. Of the 1798 tumors to which the BDs applied, 0.7% (95% CI, 0.4-1.1%) were malignant (0.3% borderline tumors, 0.1% stage I primary ovarian invasive malignancy, 0.1% stage II-IV primary ovarian invasive malignancy, 0.2% secondary metastatic tumors) (Table 3).

Among the 3107 patients with a tumor to which the BDs did not apply, the AUC for ADNEX with CA125 was 0.92 (95% CI, 0.89-0.94; 95% PI, 0.80-0.97) and for ADNEX without CA125 it was 0.91 (95% CI, 0.87-0.93; 95% PI, 0.78-0.96).

The overall AUCs for the two-step strategy were 0.95 (95% CI, 0.92-0.96; 95% PI, 0.85-0.98) when ADNEX with CA125 was used as the second step and 0.94 (95% CI, 0.92-0.96; 95% PI, 0.84-0.98) when ADNEX without CA125 was used (Figure 2). Sensitivity and specificity of the two-step strategies when using different risk cut-offs to indicate malignancy are shown in Table S7. The overall calibration curves for the two-step strategies are shown in Figure 3. Risk estimates were slightly underestimated, and heterogeneity between centers was observed (Figure S2). The summary decision curves of the two-step strategies overlapped completely with the curves showing the results when using ADNEX in all tumors (Figure 4).

The ability of the two-step strategies to discriminate between different tumor types is shown in Table S8. With two exceptions the two two-step strategies manifested similar discriminative ability: using ADNEX with CA125 as the second step instead of ADNEX without CA125 discriminated better between stage II-IV and stage I ovarian malignancy (AUC 0.81 vs 0.72) and between stage II-IV ovarian malignancy and metastases (AUC 0.76 vs 0.64). For discrimination between benign tumors and each malignant subtype, AUCs ranged from 0.91 to 0.98. Calibration of the predicted risks for the five subgroups of tumor was good for both two-step strategies, however with some underestimation of the risk of secondary metastasis (Figure S3).

Subgroup analyses

The BDs were applicable less often in post-menopausal (24%, 509/2151) than pre-menopausal women (47%, 1289/2754), and the prevalence of malignancy among tumors to which a BD applied was higher in post-menopausal than pre-menopausal women (1.0% vs 0.5%) (Table S9). The BDs were applicable less often in patients examined in oncology centers (33%, 1020/3094) than in non-oncology centers (43%, 778/1811), and the prevalence of malignancy among tumors to which a BD applied was higher in oncology centers than in non-oncology centers (0.8% vs 0.4%) (Table S10).

The discriminative ability of the two-step strategies was similar in pre- and post-menopausal women, but the two-step strategies were better calibrated in post-menopausal women (Figures S4-S9, Table S11). The discriminative ability and the calibration of the two-step strategies were similar in oncology centers and non-oncology centers (Figures S10-S15, Table S12).

Additional analyses

The results of the additional analyses are shown in Table S13-S14 and Figures S16-S21. Omitting patients with uncertain tumor outcome from the analysis slightly increased the overall prevalence of malignancy as compared to the primary analysis (22.1% vs 21.1%) but had minimal effect on

discriminative performance and calibration. Our sensitivity analysis in which we expanded the definition of uncertain outcome resulted in 1419 (29%) tumors with uncertain outcome and an AUC of the two-step strategy without CA125 of 0.93 (95% CI, 0.91-0.95; 95% PI, 0.82-0.98).

DISCUSSION

We describe the diagnostic performance of the modified benign Simple Descriptors (BDs) and of a two-step strategy using the BDs as a first step and ADNEX as a second step when applied on patients managed either surgically or conservatively. The results indicate that the BDs are applicable in almost 40% of patients with an adnexal mass, that the risk of malignancy is very low if a BD applies, and that the two-step strategy has excellent discriminative performance and is reasonably well calibrated.

The study strengths include (1) large sample size and high number of participating centers; (2) prospective ultrasound protocol with agreed ultrasound terms, definitions and measurement techniques; and (3) consecutive inclusion of patients managed either surgically or conservatively.

We acknowledge four limitations related to potential bias. First, we excluded all data from 19 centers because they did not fulfill our predefined quality criteria. This means that - like in any study - we cannot rule out selection bias on center level. We did these exclusions to obtain informative (excluding centers with limited recruitment), representative (excluding centers with non-consecutive recruitment) and reliable data (excluding centers with low quality data or centers that stopped participation). This also reduced the potential for attrition bias (exclusions on patient level). Including other centers could have resulted in higher or lower performance due to case-mix heterogeneity[31]. We do not expect our exclusion of centers to have resulted in an overestimation of diagnostic performance, because we do not expect the quality of the ultrasound examinations to be lower in the excluded than in the included centers. In a study on the same dataset, a sensitivity analysis using immediately operated patients from all 36 centers resulted in the same AUC for ADNEX as the primary analysis based on 17 centers[17]. Second, the tumor outcome was based on multiple reference standards ('differential verification'), and for a small group of patients tumor outcome could not be determined due to conflicting information or insufficient follow-up ('partial verification')[29]. We addressed the potentially optimistic bias from partial and differential verification by using multiple imputation and a sensitivity analysis[27]. Model performance changed very little depending on the definition of uncertainty that we used. Third, 53% of the patients had a missing value for CA125. This affects the performance of ADNEX with CA125 but not that of the BDs or ADNEX without CA125. To deal with missing CA125 values, we used multiple imputation, which is the recommended approach to reduce bias due to missing values[18,30]. Excluding cases with missing CA-125 is likely to bias AUCs downwards, because missing values are most common among tumors judged to be benign on ultrasound (Table S4) [32]. The high number of missing values adds uncertainty to the results, and imputing multiple times acknowledges this as reflected in wider confidence intervals around performance estimates. Fourth, there was no blinding of examiners to previous information about the patient or of pathologists to clinical information. Imposing such blinding would be unethical and unrealistic. Lack of blinding may induce information bias when assessing predictors and detection bias when determining outcome based on histology. Information bias

may lead to overestimation of performance (even though, importantly, the outcome was unknown at the inclusion scan when the predictors were assessed). We consider detection bias to be limited. Pathologists are unlikely to be influenced by preoperative ultrasound findings. This assumption is supported by findings in the IOTA phase 1 study, in which the results of central pathology review (blinded to clinical information) were highly similar to local results[8]. In summary, we used several recommended approaches to reduce bias. Information bias - an unavoidable clinical reality - may nevertheless have optimistically biased performance.

Our study is the first to evaluate the performance of the BDs, the first to suggest and evaluate the performance of a two-step strategy using the BDs as a first step and ADNEX as a second step, and to do this in patients managed either surgically or conservatively. Three studies have validated a three-step strategy using both the benign and malignant original Simple Descriptors as a first step, followed by the IOTA Simple Rules[10] as a second step, and by subjective assessment by an expert as a final third step. All three studies included patients managed either surgically or conservatively and showed the three-step strategy to have excellent ability to discriminate between benign and malignant adnexal masses[33–35].We believe that our two-step strategy has advantages over the three-step strategy: all tumors can be classified by one single ultrasound examiner, a risk of malignancy is assigned to all tumors as well as a likelihood estimate of type of malignancy.

It is reassuring that the discriminative ability of ADNEX when applied on tumors to which the BDs do not apply was almost as good (AUC 0.92 and 0.91) as when ADNEX was applied on all tumors. When applied on all 4905 tumors, the AUC for ADNEX both with and without CA125 was 0.9417. Moreover, the discriminative performance of the two-step strategies (AUC 0.95 and 0.94) was similar to that of using ADNEX on all 4905 masses, and the clinical utility of the two-step strategies was the same as that of applying ADNEX on all masses (Figure 4). This shows that using ADNEX on all masses has no advantage over using the two-step strategy.

Two issues require further research. First, evaluating ultrasound images is affected by the level of experience of the examiner. The IOTA5 study involves mainly level II and III examiners. Even though previous studies have suggested that ADNEX works well also in the hands of less experienced examiners[36] the role of experience should be investigated more explicitly. A large multicenter study, in which examiner experience is quantified before patient recruitment, could elucidate whether and how experience affects discrimination and calibration performance. Second, we observed heterogeneity between centers regarding discrimination and calibration for all models validated on IOTA phase 5 data[17]. When more data becomes available, it is important to study possible reasons for this heterogeneity.

The two-step strategy lends itself very well to clinical use. A large proportion of adnexal masses can be classified by the BDs as having a very low risk of malignancy without computer support. For the

remaining masses an estimate of risk of malignancy and type of malignancy can be obtained by using the ADNEX model. An ADNEX calculator is available online and as an application for smartphones (https://iotagroup.org/iota-models-software/adnex-risk-model). It is also embedded in many ultrasound-machines. This facilitates its use in clinical practice. The two-step strategy can be used for patient counselling to individualize management. It could also be used to stratify patients into risk groups, e.g. into the Ovarian-Adnexal reporting & Data system (O-RADS) risk groups[37]. Risk stratification can facilitate choosing the optimal management for patients with adnexal masses[38].

DISCLOSURE

Funding

The IOTA5 study is supported by the Research Foundation-Flanders (FWO) projects G097322N /G049312N/G0B4716N/12F3114N, and Internal Funds KU Leuven (projects C24/15/037 and C24M/20/064). Dirk Timmerman is a senior clinical investigator of FWO, Laure Wynants is a post-doctoral fellow of FWO, and Wouter Froyman was a clinical fellow of FWO. Chiara Landolfo was supported by Linbury Trust Grant LIN 2600. Tom Bourne is supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial College Healthcare National Health Service (NHS) Trust and Imperial College London. The views expressed are those of the authors and not necessarily those of the NHS, NIHR or Department of Health. Lil Valentin is supported by the Swedish Research Council (grant K2014-99X-22475-01-3, Dnr 2013-02282), funds administered by Malmö University Hospital and Skåne University Hospital, Allmänna Sjukhusets i Malmö Stiftelse för bekämpande av cancer (the Malmö General Hospital Foundation for fighting against cancer), and two Swedish governmental grants (Avtal om läkarutbildning och forskning (ALF)-medel and Landstingsfinansierad Regional Forskning).

Role of the funding source

The funders played no role in study design, data collection, data analysis, data interpretation, writing of the report, or decision to submit the manuscript for publication. The guarantors had full access to all the data in the study, take responsibility for the integrity of the data and the accuracy of the data analysis, and had final responsibility for the decision to submit for publication.

REFERENCES

1. Querleu D, Planchamp F, Chiva L, Fotopoulou C, Barton D, Cibula D, Aletti G, Carinelli S, Creutzberg C, Davidson B, Harter P, Lundvall L, Marth C, Morice P, Rafii A, Ray-Coquard I, Rockall A, Sessa C, Van Der Zee A, Vergote I, Du Bois A. European society of gynaecologic oncology quality indicators for advanced ovarian cancer surgery. *Int J Gynecol Cancer* 2016; **26**: 1354–1363.

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**: 286–290.

3. Engelen MJA, Kos HE, Willemse PHB, Aalders JG, De Vries EGE, Schaapveld M, Otter R, Van Der Zee AGJ. Surgery by consultant gynecologic oncologists improves survival in patients with ovarian carcinoma. *Cancer* 2006; **106**: 589–598.

4. Bristow RE, Chang J, Ziogas A, Anton-Culver H. Adherence to treatment guidelines for ovarian cancer as a measure of quality care. *Obstet Gynecol* 2013; **121**: 1226–1234.

5. Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas J. 90346536 A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Maturitas* 1991; **13**: 177.

6. Testa A, Kaijser J, Wynants L, Fischerova D, Van Holsbeke C, Franchi D, Savelli L, Epstein E, Czekierdowski a., Guerriero S, Fruscio R, Leone FPG, Vergote I, Bourne T, Valentin L, Van Calster B, Timmerman D. Strategies to diagnose ovarian cancer: New evidence from phase 3 of the multicentre international IOTA study. *Br J Cancer* 2014; **111**: 680–688.

7. Van Calster B, Van Hoorde K, Valentin L, Testa a. C, Fischerova D, Van Holsbeke C, Savelli L, Franchi D, Epstein E, Kaijser J, Van Belle V, Czekierdowski a., Guerriero S, Fruscio R, Lanzani C, Scala F, Bourne T, Timmerman D. 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.

8. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L. 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**: 8794–8801.

9. Timmerman D, Van Calster B, Testa A, Savelli L, Fischerova D, Froyman W, Wynants L, Van Holsbeke C, Epstein E, Franchi D, Kaijser J, Czekierdowski A, Guerriero S, Fruscio R, Leone FPG, Rossi A, Landolfo C, Vergote I, Bourne T, Valentin L. 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**: 424–437.

10. Timmerman D, Testa A. C, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, Paladini D, Van Calster B, Vergote I, Van Huffel S, Valentin L. Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol* 2008; **31**: 681–690.

11. Westwood M, Ramaekers B, Lang S, Grimm S, Deshpande S, de Kock S, Armstrong N, Joore M, Kleijnen J. 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**: V-264.

12. Meys EMJ, Kaijser J, Kruitwagen RFPM, Slangen BFM, Van Calster B, Aertgeerts B, Verbakel JY, Timmerman D, Van Gorp T. Subjective assessment versus ultrasound models to diagnose ovarian cancer: A systematic review and meta-analysis. *Eur J Cancer* 2016; **58**: 17–29.

13. Kaijser J, Sayasneh A, Van hoorde K, Ghaem-maghami S, Bourne T, Timmerman D, Van Calster B. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: A systematic review and meta-analysis. *Hum Reprod Update* 2014; **20**: 449–462.

14. Sayasneh A, Wynants L, Preisler J, Kaijser J, Johnson S, Stalder C, Husicka R, Abdallah Y, Raslan F, Drought A, Smith AA, Ghaem-Maghami S, Epstein E, Van Calster B, Timmerman D, Bourne T. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer* 2013; **108**: 2448–2454.

15. Ameye L, Timmerman D, Valentin L, Paladini D, Zhang J, Van Holsbeke C, Lissoni a. a., Savelli L, Veldman J, Testa a. C, Amant F, Van Huffel S, Bourne T. Clinically oriented three-step strategy for assessment of adnexal pathology. *Ultrasound Obstet Gynecol* 2012; **40**: 582–591.

16. Froyman W, Landolfo C, De Cock B, Wynants L, Sladkevicius P, Testa AC, Van Holsbeke C, Domali E, Fruscio R, Epstein E, dos Santos Bernardo MJ, Franchi D, Kudla MJ, Chiappa V, Alcazar JL, Leone FPG, Buonomo F, Hochberg L, Coccia ME, Guerriero S, Deo N, Jokubkiene L, Kaijser J, Coosemans A, Vergote I, Verbakel JY, Bourne T, Van Calster B, Valentin L, Timmerman D. 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**: 448–458.

17. Van Calster B, Valentin L, Froyman W, Landolfo C, Ceusters J, Testa AC, Wynants L, Sladkevicius P, Van Holsbeke C, Domali E, Fruscio R, Epstein E, Franchi D, Kudla MJ, Chiappa V, Alcazar JL, Leone FPG, Buonomo F, Coccia ME, Guerriero S, Deo N, Jokubkiene L, Savelli L, Fischerová D, Czekierdowski A, Kaijser J, Coosemans A, Scambia G, Vergote I, Bourne T, Timmerman D. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. *BMJ* 2020; **370**: m2614.

18. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–W73.

19. Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. 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**: 500–505.

20. Education and Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) Minimum training recommendations for the practice of medical ultrasound. *Ultraschall Med* 2006; **27**: 79–105.

21. Valentin L. Use of morphology to characterize and manage common adnexal masses. *Best Pract Res Clin Obstet Gynaecol* 2004; **18**: 71–89.

22. Prat J, Committee F. International Journal of Gynecology and Obstetrics Staging classification for cancer of the ovary, fallopian tube, and peritoneum. *Int J Gynecol Obstet* 2014; **124**: 1–5.

23. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, Collins GS, MacAskill P, Moons KGM, Vickers AJ. Calibration: The Achilles heel of predictive analytics. *BMC Med* 2019; **17**: 1–7.

24. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**: 3–7.

25. 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**: 761–770.

26. Van Hoorde K, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW, Van Calster B. Assessing calibration of multinomial risk prediction models. *Stat Med* 2014; **33**: 2585–2596.

27. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353**: i3140–i3140.

28. Snell KIE, Ensor J, Debray TPA, Moons KGM, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2018; **27**: 3505–3522.

29. De Groot JAH, Bossuyt PMM, Reitsma JB, Rutjes AWS, Dendukuri N, Janssen KJM, Moons KGM. Verification problems in diagnostic accuracy studies: Consequences and solutions. *BMJ* 2011; **343**: 1–9.

30. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Ann Intern Med* 2019; **170**: W1–W33.

31. Steyerberg EW. Clinical Prediction Models: a Practical Approach to Development, Validation, and Updating (2nd ed). Cham, Springer: 2019.

32. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ*. 2009; **339**: 157–160.

33. Peces Rama A, Llanos Llanos MC, Sánchez Ferrer ML, Alcázar Zambrano JL, Martínez Mendoza A, Nieto Díaz A. Simple descriptors and simple rules of the International Ovarian Tumor Analysis (IOTA) Group: A prospective study of combined use for the description of adnexal masses. *Eur J Obstet Gynecol Reprod Biol* 2015; **195**: 7–11.

34. Hidalgo JJ, Ros F, Aubá M, Errasti T, Olartecoechea B, Ruiz-Zambrana, Alcázar JL. Prospective external validation of IOTA three-step strategy for characterizing and classifying adnexal masses and retrospective assessment of alternative two-step strategy using simple-rules risk. *Ultrasound Obstet Gynecol* 2019; **53**: 693–700.

35. Alcázar JL, Pascual MA, Graupera B, Aubá M, Errasti T, Olartecoechea B, Ruiz-Zambrana A, Hereter L, Ajossa S, Guerriero S. External validation of IOTA simple descriptors and simple rules for classifying adnexal masses. *Ultrasound Obstet Gynecol* 2016; **48**: 397–402.

36. Sayasneh A, Ferrara L, De Cock B, Saso S, Al-Memar M, Johnson S, Kaijser J, Carvalho J, Husicka R, Smith A, Stalder C, Blanco MC, Ettore G, Van Calster B, Timmerman D, Bourne T. Evaluating the risk of ovarian cancer before surgery using the ADNEX model: A multicentre external validation study. *Br J Cancer* 2016; **115**: 542–548.

37. Andreotti RF, Timmerman D, Strachowski LM, Froyman W, Benacerraf BR, Bennett GL, Bourne T, Brown DL, Coleman BG, Frates MC, Goldstein SR, Hamper UM, Horrow MM, Hernanz-Schulman M, Reinhold C, Rose SL, Whitcomb BP, Wolfman WL, Glanc P. O-RADS US risk stratification and management system: A consensus guideline from the ACR ovarian-Adnexal Reporting and Data System committee. *Radiology* 2020; **294**: 168–185.

38. Timmerman D, Planchamp F, Bourne T, Landolfo C, Du Bois A, Chiva L, Cibula D, Concin N, Fischerova D, Froyman W, Gallardo Madueño G, Lemley B, Loft A, Mereu L, Morice P, Querleu

D, Testa AC, Vergote I, Vandecaveye V, Scambia G, Fotopoulou C. ESGO/ISUOG/IOTA/ESGE Consensus Statement on pre-operative diagnosis of ovarian tumors. *Int J Gynecol Cancer* 2021; **31**: 961–982.

FIGURE LEGENDS

Figure 1. International Ovarian Tumor Analysis (IOTA) modified benign Simple Descriptors (Benign Descriptors) illustrated by ultrasound images.

Figure 2. Forest plot with center-specific AUCs of the two-step strategies and results of meta-analysis (n=4905). ”Other” includes the following small non-oncology centers with low prevalence of malignancy: London (UK), Nottingham (UK), Milan 3 (Italy), and Florence (Italy). AUC, area under the receiver operating characteristic curve; ADNEX, Assessment of Different NEoplasias in the adneXa; CI, confidence interval.

Figure 3. Overall calibration curves of the two-step strategies using ADNEX with CA125 or ADNEX without CA125 as a second step. BD, Benign Descriptors; ADNEX, Assessment of Different NEoplasias in the adneXa; Intercept, calibration intercept; Slope, calibration slope; CI, confidence interval.

Figure 4. Overall decision curves for the ADNEX models and for the two-step strategies (n=4905). ADNEX, Assessment of Different NEoplasias in the adneXa; BD, Benign Descriptors

Table 1. Definition of tumor outcome based on histology or clinical information (also published in BMJ[17]).

Outcome	Scenario	N
Benign	B1: Surgery, benign histology	2065
	B2: No surgery, no spontaneous resolution, last visit ≥10 months, SA at every visit up to 14 months was 'probably benign' or 'certainly benign'	911
	B3: Spontaneous resolution	465
Malignant	M1: Surgery within 120 days, malignant histology	956
	M2: Surgery after 120 days, malignant histology, SA at every visit up to surgery was 'probably borderline/malignant' or 'certainly borderline/malignant'	18*
	M3: No surgery, no spontaneous resolution, last visit ≥10 months, SA at every visit up to 14 months was 'probably borderline/malignant' or 'certainly borderline/malignant'	4†
Uncertain	U1: Surgery after 120 days, malignant histology, SA not 'probably borderline/malignant' or 'certainly borderline/malignant' at every visit up to surgery	19*
	U2: No surgery, no spontaneous resolution, last visit ≥10 months, SA was 'uncertain' or was inconsistent across visits up to 14 months	35
	U3: No surgery, no spontaneous resolution, last follow-up visit was before 10 months (due to death, withdrawal from study, or lost to follow-up)	123
	U4: No information since the inclusion visit	309

B=benign; M=malignant; U=uncertain at inclusion; SA=subjective assessment.

* In line with previous publications,7 we used 120 days as the maximum interval between inclusion and surgery. When surgery was done more than 120 days after inclusion and histology was malignant, we recognize the possibility that the tumor was benign at inclusion but underwent malignant transformation. In these cases, we relied on subjective assessment at inclusion and on follow-up scans to decide whether to label the outcome as 'malignant' or as 'uncertain'.

†In these cases, the type of malignancy could not be determined. Type of malignancy was treated as a missing value and imputed (Appendix S4).

Table 2. Descriptive statistics for patients and tumors included (n = 4905).

Variable	Median (IQR) or n (%)
Patient age at recruitment (years)	
Median and interquartile range	48 (IQR 36 – 62)
Range	18 – 98
Postmenopausal*	2151 (44%)
Gynecological symptoms during the year preceding inclusion, n (%)	2565 (52%)
Serum CA125 (U/mL)	
Median and interquartile range	25 (IQR 12 – 107)
Range	1 – 57900
Missing values	2620 (53%)
Bilateral masses	829 (17%)
Ascites	285 (6%)
Largest diameter of lesion (mm)	
Median and interquartile range	55 (38 – 83)
Range	7 – 751
Tumor type using IOTA terminology	
Unilocular	2140 (44%)
Unilocular-solid	396 (8%)
Multilocular	1011 (21%)
Multilocular-solid	649 (13%)
Solid	689 (14%)
Not possible to classify	20 (0.4 %)
Presence of solid components	1734 (35%)
Largest diameter of largest solid component†	
Median and interquartile range	41 (19 – 68)
Range	3 – 751
Number of papillary projections	
None	4335 (88%)
1	282 (6%)
2	85 (2%)
3	49 (1%)
More than 3	154 (3%)
More than 10 cyst locules	368 (8%)
Irregular internal cyst walls	1502 (31%)
Echogenicity of cyst fluid	
Anechoic	1852 (38%)
Low level	778 (16%)
Ground glass	793 (16%)
Mixed	680 (14%)
Hemorrhagic	113 (2%)
Not applicable	689 (14%)
Color score, n (%)	
1: no blood flow	2031 (41%)
2: minimal blood flow	1336 (27%)
3: moderate blood flow	1099 (22%)
4: very strong flow	439 (9%)
Ultrasound examiner's subjective impression	
Certainly benign	2488 (51%)
Probably benign	1066 (22%)
Uncertain	367 (7%)
Probably malignant	392 (8%)
Certainly malignant	592 (12%)

* Menopausal status at recruitment. If menopausal status was uncertain (e.g. because of hysterectomy), we classified patients aged 50 years or older as postmenopausal.

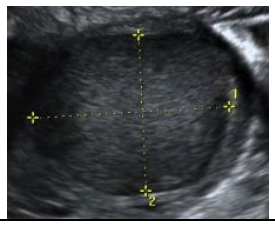
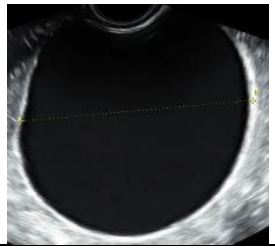
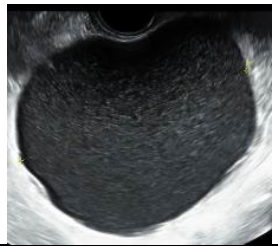
†Only for tumors with a solid component

Table 3. Prevalence of each tumor subtype among masses to which a Benign Descriptor applied (n=1798; pooled data).

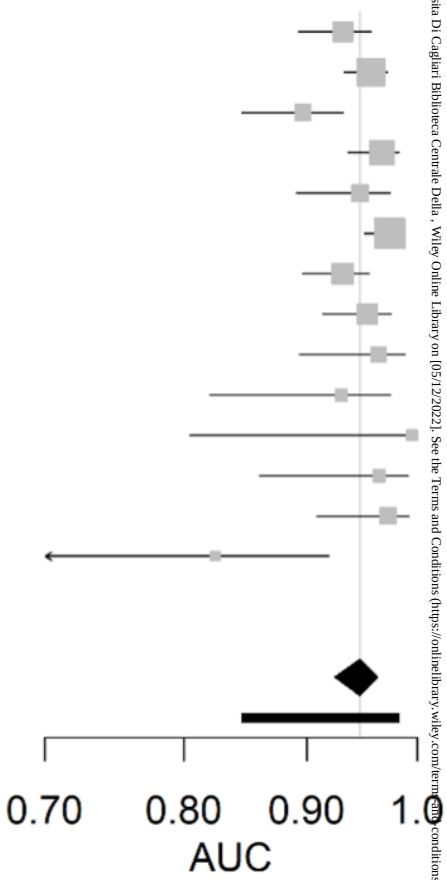
		Tumor subtype				
Benign Descriptor	**n**	**Benign n (%)**	**Borderline n (%)**	**Stage I invasive n (%)**	**Stage II – IV invasive n (%)**	**Secondary metastatic n (%)**
Any descriptor	1798	1786.3 (99.3)	5.0 (0.3)	1.5 (0.1)	1.6 (0.1)	3.6 (0.2)
Descriptor 1	514	511.9 (99.6)	1.6 (0.3)	0.3 (0.1)	0.2 (<0.1)	0.1 (<0.1)
Descriptor 2	185	>184.9 (>99.9)	<0.1 (<0.1)	0.0 (0.0)	0.0 (0.0)	<0.1 (<0.1)
Descriptor 3	692	689.2 (99.6)	1.3 (0.2)	0.1 (<0.1)	1.2 (0.2)	0.2 (<0.1)
Descriptor 4	407	400.2 (98.3)	2.1 (0.5)	1.2 (0.3)	0.2 (0.1)	3.3 (0.8)

Percentages are calculated per row. Uncertain tumor outcomes have been multiply imputed. The table shows results averaged over imputed datasets, hence the decimals for the number of tumors of each type. There are no decimals for the overall n because the data needed for the Benign Descriptors did not have missing data.

Figure 1. International Ovarian Tumor Analysis (IOTA) modified benign Simple Descriptors (Benign Descriptors) illustrated by ultrasound images.

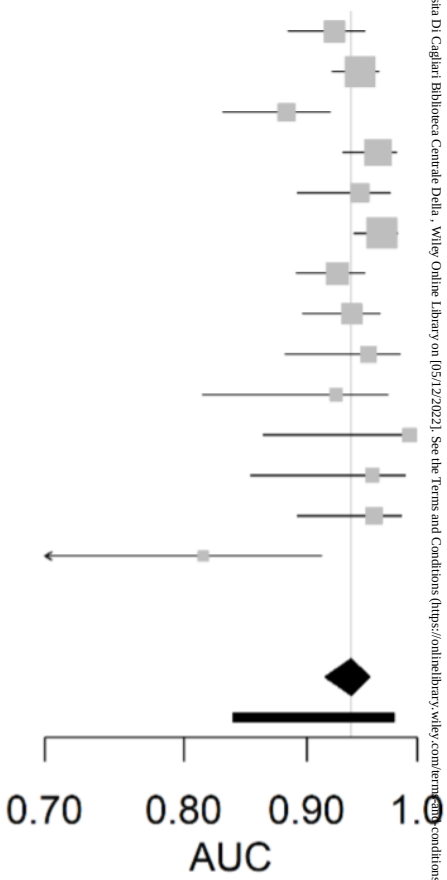
	Benign descriptor 1: Unilocular cyst with ground glass echogenicity and largest diameter <10 cm in a premenopausal woman (suggestive of endometrioma)
	Benign descriptor 2: Unilocular cyst with mixed echogenicity, acoustic shadows and largest diameter <10 cm in a premenopausal woman (suggestive of benign cystic teratoma)
	Benign descriptor 3: Unilocular cyst with anechoic cyst fluid, smooth internal walls and largest diameter <10 cm (suggestive of simple cyst or cystadenoma)
	Benign descriptor 4: All other unilocular cysts with smooth internal walls and largest diameter <10 cm

Center	AUC (95% CI)	N
Malmö (Sweden)	0.93 (0.89 to 0.96)	794
Rome (Italy)	0.96 (0.93 to 0.97)	681
Athens (Greece)	0.90 (0.85 to 0.93)	567
Leuven (Belgium)	0.97 (0.94 to 0.98)	501
Genk (Belgium)	0.95 (0.89 to 0.97)	406
Milan 1 (Italy)	0.97 (0.95 to 0.99)	367
Stockholm (Sweden)	0.93 (0.90 to 0.95)	363
Monza (Italy)	0.95 (0.91 to 0.98)	267
Cagliari (Italy)	0.96 (0.89 to 0.99)	166
Katowice (Poland)	0.93 (0.82 to 0.97)	139
Pamplona (Spain)	1.00 (0.80 to 1.00)	111
Trieste (Italy)	0.96 (0.86 to 0.99)	111
Milan 2 (Italy)	0.97 (0.91 to 0.99)	98
Other	0.82 (0.66 to 0.92)	334
Meta-analysis		
AUC (95% CI)	**0.95 (0.92 to 0.96)**	**4905**
95% Prediction interval	**(0.85 to 0.98)**	

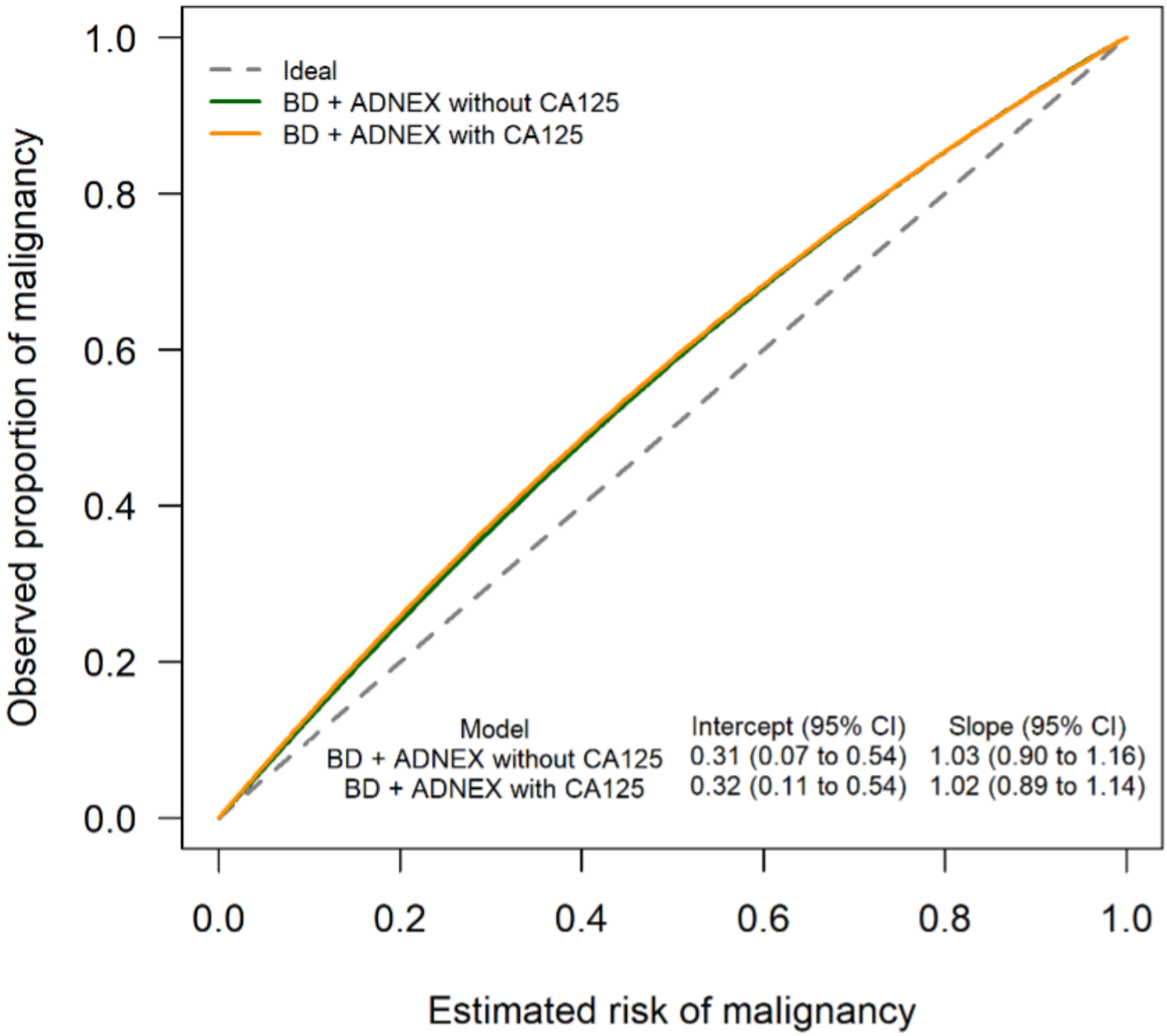


UOG_26080_Figure 2 A.png

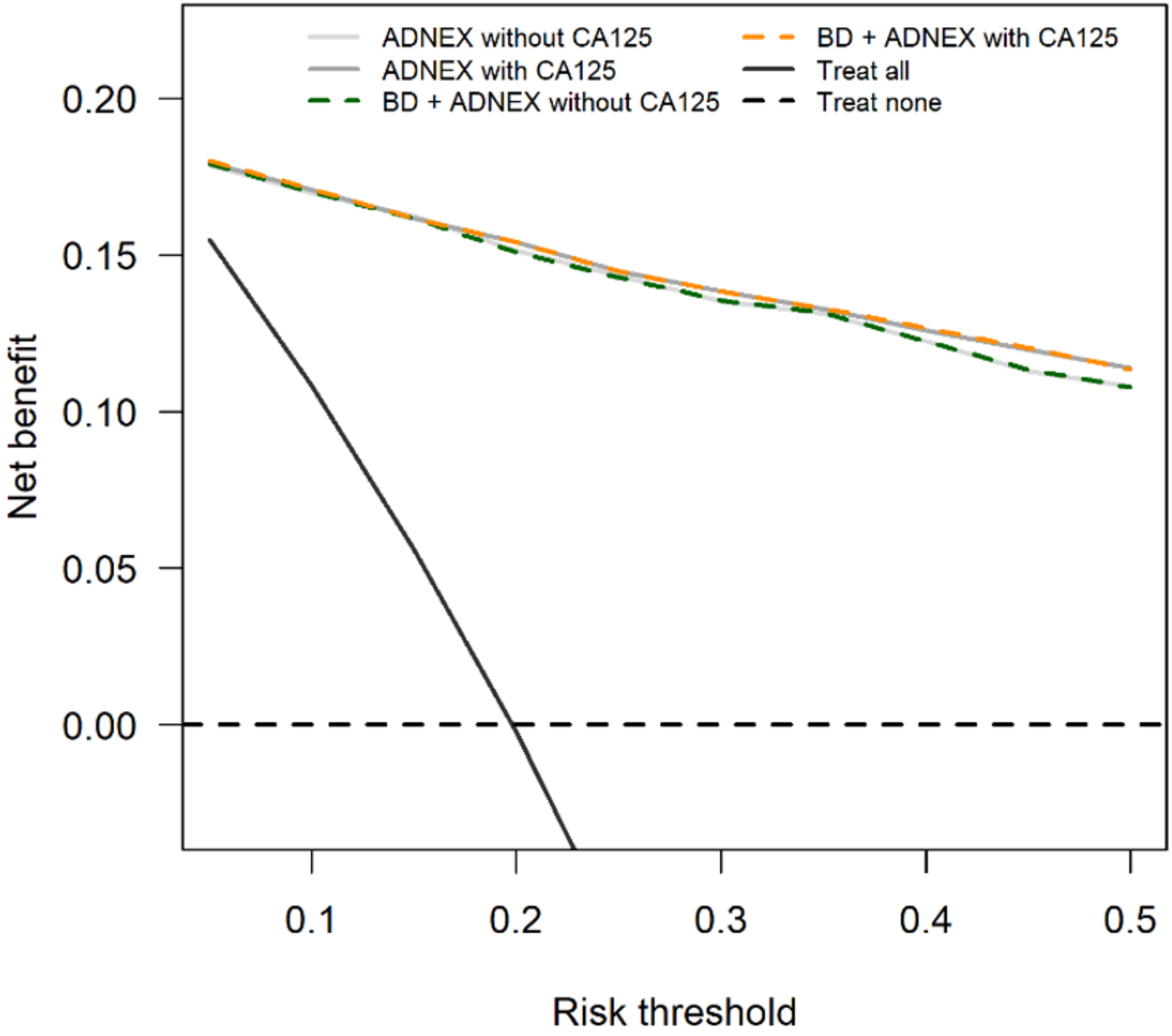
Center	AUC (95% CI)	N
Malmö (Sweden)	0.92 (0.88 to 0.95)	794
Rome (Italy)	0.95 (0.92 to 0.96)	681
Athens (Greece)	0.88 (0.83 to 0.92)	567
Leuven (Belgium)	0.96 (0.93 to 0.98)	501
Genk (Belgium)	0.95 (0.89 to 0.97)	406
Milan 1 (Italy)	0.97 (0.94 to 0.98)	367
Stockholm (Sweden)	0.93 (0.89 to 0.95)	363
Monza (Italy)	0.94 (0.90 to 0.97)	267
Cagliari (Italy)	0.95 (0.88 to 0.98)	166
Katowice (Poland)	0.93 (0.81 to 0.97)	139
Pamplona (Spain)	0.99 (0.86 to 1.00)	111
Trieste (Italy)	0.96 (0.85 to 0.99)	111
Milan 2 (Italy)	0.96 (0.89 to 0.99)	98
Other	0.81 (0.65 to 0.91)	334
Meta-analysis		
AUC (95% CI)	**0.94 (0.92 to 0.96)**	**4905**
95% Prediction interval	**(0.84 to 0.98)**	



UOG_26080_Figure 2 B.png



UOG_26080_Figure 3.png



UOG_26080_Figure 4.png