FULL PAPER

BJC

British Journal of Cancer (2014) 111, 680–688 | doi: 10.1038/bjc.2014.333

Keywords: ovarian cancer; diagnosis; meta-analysis

Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study

A Testa[1,17], J Kaijser[2,3,17], L Wynants[4,5], D Fischerova[6], C Van Holsbeke[7], D Franchi[8], L Savelli[9], E Epstein[10], A Czekierdowski[11], S Guerriero[12], R Fruscio[13], F P G Leone[14], I Vergote[3], T Bourne[2,3,15], L Valentin[16], B Van Calster[2] and D Timmerman*[,2,3]

[1]*Department of Gynaecologic Oncology, Catholic University of the Sacred Heart, Largo Francesco Vito 8, Rome 00165, Italy;* [2]*KU Leuven Department of Development and Regeneration, Herestraat 49 Box 7003, 3000 Leuven, Belgium;* [3]*Department of Obstetrics and Gynaecology and Leuven Cancer Institute, University Hospitals Leuven, Herestraat 49 Box 7003, 3000 Leuven, Belgium;* [4]*KU Leuven Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium;* [5]*iMinds Future Health Department, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium;* [6]*Department of Obstetrics and Gynaecology, Gynaecological Oncology Center, Charles University, Apolinarska 18, 12000 Prague, Czech Republic;* [7]*Department of Obstetrics and Gynaecology, Ziekenhuis Oost Limburg, Schiepse Bos 6, 3600 Genk, Belgium;* [8]*Preventive Gynaecology Unit, Division of Gynaecology, European Institute of Oncology, Via Ripamonti 435, Milan 20141, Italy;* [9]*Gynaecology and Reproductive Medicine Unit, S. Orsola-Malpighi Hospital, University of Bologna, Via Albertoni 15, Bologna 40138, Italy;* [10]*Department of Obstetrics and Gynaecology, Karolinska University Hospital, Solna, SE-17176 Stockholm, Sweden;* [11]*First Department of Gynecologic Oncology and Gynecology, Medical University of Lublin, ul. Staszica 16, 20-081 Lublin, Poland;* [12]*Department of Obstetrics and Gynaecology, Azienda Ospedaliero Universitaria di Cagliari, Strada Statale 554 Monserrato, Cagliari 09045, Italy;* [13]*Department of Obstetrics and Gynaecology, San Gerardo Hospital, University of Milan-Bicocca, Via Pergolesi, 33, 20052 Monza, Italy;* [14]*Department of Obstetrics and Gynecology, Clinical Sciences Institute L. Sacco, University of Milan, Via G.B. Grassi 74, 20157 Milan, Italy;* [15]*Department of Obstetrics and Gynaecology, Queen Charlotte's and Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, UK and* [16]*Department of Obstetrics and Gynaecology, Skåne University Hospital Malmö, Lund University, Södra Förstadsgatan, 20502 Malmö, Sweden*

Background: To compare different ultrasound-based international ovarian tumour analysis (IOTA) strategies and risk of malignancy index (RMI) for ovarian cancer diagnosis using a meta-analysis approach of centre-specific data from IOTA3.

Methods: This prospective multicentre diagnostic accuracy study included 2403 patients with 1423 benign and 980 malignant adnexal masses from 2009 until 2012. All patients underwent standardised transvaginal ultrasonography. Test performance of RMI, subjective assessment (SA) of ultrasound findings, two IOTA risk models (LR1 and LR2), and strategies involving combinations of IOTA simple rules (SRs), simple descriptors (SDs) and LR2 with and without SA was estimated using a meta-analysis approach. Reference standard was histology after surgery.

Results: The areas under the receiver operator characteristic curves of LR1, LR2, SA and RMI were 0.930 (0.917–0.942), 0.918 (0.905–0.930), 0.914 (0.886–0.936) and 0.875 (0.853–0.894). Diagnostic one-step and two-step strategies using LR1, LR2, SR and SD achieved summary estimates for sensitivity 90–96%, specificity 74–79% and diagnostic odds ratio (DOR) 32.8–50.5. Adding SA when IOTA methods yielded equivocal results improved performance (DOR 57.6–75.7). Risk of Malignancy Index had sensitivity 67%, specificity 91% and DOR 17.5.

Conclusions: This study shows all IOTA strategies had excellent diagnostic performance in comparison with RMI. The IOTA strategy chosen may be determined by clinical preference.

*Correspondence: Professor D Timmerman; E-mail: dirk.timmerman@uzleuven.be
[17]The first two authors should be regarded as joint First Authors.

Received 9 February 2014; revised 12 May 2014; accepted 14 May 2014; published online 17 June 2014

Providing care within highly specialised multidisciplinary services has a clear survival benefit for patients with ovarian cancer (Woo *et al*, 2012). Although such centralised care is recommended in many developed countries, a large proportion of ovarian cancer patients remain treated by general surgeons and physicians (Verleye *et al*, 2010). Several factors probably contribute to this failure to refer for specialist care, but the lack of effective preoperative strategies to evaluate ovarian tumours is certainly one of the most important factors (Miller and Ueland, 2012). Reports from the international ovarian tumour analysis (IOTA) multicentre studies phase 1, 1b, 2 and 4 (Timmerman *et al*, 2005, 2007, 2008, 2010a, b; Van Holsbeke *et al*, 2009, 2012; Ameye *et al*, 2012; Kaijser *et al*, 2013; Sayasneh *et al*, 2013a, b) have demonstrated that IOTA ultrasound-based approaches to characterise adnexal masses in the hands of physicians and sonographers with varying levels of experience outperform other established strategies such as use of individual biomarkers (serum CA-125), the Risk of Malignancy Index (RMI; Jacobs *et al*, 1990) or Risk of Ovarian Malignancy Algorithm (Moore *et al*, 2009), for the classification of ovarian pathology. Nevertheless, there is a paucity of comprehensive prospective studies comparing different diagnostic strategies for ovarian cancer diagnosis on the same study population. Such studies are of pivotal importance for assessing diagnostic test accuracy. In the most recently published meta-analysis, only a small number of the studies included validated different diagnostic tests for ovarian cancer on the same data set (Kaijser *et al*, 2014).

The primary aim of this study, the IOTA phase 3 study, was to compare the test performance of various IOTA diagnostic strategies and RMI on prospectively collected data from a large number of patients and centres.

MATERIALS AND METHODS

Study design. This was a multicentre cross-sectional diagnostic accuracy study with prospective data collection. Patients were recruited between October 2009 and May 2012, in 18 centres in six countries (Sweden, Belgium, Italy, Poland, Spain and Czech Republic). These centres were either oncology referral centres (i.e., tertiary centres for the treatment of women with gynaecological malignancy) or general hospitals and units with a special interest in gynaecological ultrasound. The centres and type of centres in IOTA3 are listed in a Supplementary Appendix. All centres except three (*SSW*, *BSP* and *FIT*) had participated in at least one of the previous IOTA studies (1, 1b or 2). Ethics approval was obtained by the ethics committee of the University Hospitals Leuven as main investigating centre (B32220095331/S51375) as well of the local committees of all contributing centres to IOTA3.

Inclusion criteria. Patients were eligible if they presented with at least one adnexal mass (ovarian, para-ovarian or tubal), underwent transvaginal ultrasound examination by a principal investigator at one of the participating centres and were then selected for surgical intervention by the managing clinician. Patients were examined following the research protocol if they gave informed consent. If more than one adnexal mass was detected, the mass with the most complex ultrasound morphology was denoted by the ultrasound examiner as the dominant mass, that is, the one to be used for statistical analysis. If both masses had similar morphology, the largest one or the one most easily accessible by ultrasound was denoted dominant.

Exclusion criteria. Exclusion criteria were surgical removal of the mass >120 days after the ultrasound examination, pregnancy at scan and data inconsistencies that persisted after final manual data checks.

Data collection. A dedicated, secure electronic data-collection system was developed for the study (IOTA3 Study Screen; Astraia Software, Munich, Germany). Patients automatically received a unique identifier. Data security was ensured by encrypting all data communication. Data integrity and completeness were ensured by client-side checks in the system supplied by Astraia and final data cleaning by a group of biostatisticians and expert ultrasound examiners in Leuven, Belgium.

Ultrasound examination. All included patients underwent standardised transvaginal ultrasonography by examiners experienced in gynaecologic ultrasound (level III) (Education, Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology, 2006). High-end ultrasound systems, the same or similar to those in IOTA phase 1 and 2, were used. Grey scale and colour Doppler ultrasound imaging was used to obtain information on >40 morphological and blood-flow variables to characterise each adnexal mass. Details on the ultrasound examination technique and the IOTA terms and definitions used to describe adnexal pathology have been published elsewhere (Timmerman *et al*, 2000). After completing the ultrasound examination, the ultrasound examiner classified each mass as benign or malignant on the basis of his/her subjective assessment (SA) of grey scale and colour or power Doppler ultrasound findings. Each mass was classified as certainly benign, probably benign, uncertain but most probably benign, uncertain but most probably malignant, probably malignant or certainly malignant. The ultrasound information was recorded prospectively in the electronic data-collection system, was locked at the time of the examination and could not be changed thereafter. Predictions of all diagnostic strategies under consideration (except SA) were obtained centrally after the conclusion of the study, and had no role in the decision-making process. Decision-making regarding surgery for adnexal tumours was based on clinical information (such as symptoms, age, operative risk, coexisting disease, etc.) and on the clinical ultrasound report. The clinical ultrasound report was written on the basis of the results of SA.

Serum tumour marker. Centres were encouraged to measure the level of serum CA-125 from all patients, but the availability of this biochemical end point was not a requirement for recruitment into the study.

Diagnostic strategies. The methods and strategies prospectively compared on the IOTA3 data set are SA, two IOTA logistic regression models, that is, LR1 and LR2 (Timmerman *et al*, 2005), the IOTA Simple Rules (SRs; Timmerman *et al*, 2008), the IOTA Simple Descriptors (SDs; Ameye *et al*, 2012) and various combinations of these, and the RMI. The IOTA methods are briefly described in Table 1. Details can be found in the literature (Timmerman *et al*, 2005, 2008; Ameye *et al*, 2012).

We evaluated five one-stage strategies, five two-stage strategies and two three-stage strategies.

The one-stage strategies are: the use in all patients of either LR1, LR2, SA, RMI or SRs (classifying all tumours where the SRs yield an inconclusive result as malignant).

The two-step strategies are: SRs as a first stage test and SA for tumours in which SRs yield an inconclusive result; LR2 as a first stage test and SA for tumours in which LR2 yields a predicted risk of malignancy of $\geqslant 5\%$ but $<25\%$ (risk of malignancy of $\geqslant 5\%$ but $<25\%$ arbitrarily being taken to represent an equivocal result); SDs as a first stage test, SRs for tumours unclassifiable by the SDs and tumours unclassifiable by the SRs classified as malignant; SDs as a first stage test and LR2 for those tumours where the SDs are not applicable; SDs as a first stage test and SA for those tumours in which the SDs are not applicable.

The three-step strategies are: SDs as a first stage test, SRs for tumours in which the SDs are not applicable and SA for masses in

Table 1. Description of the IOTA methods evaluated in the IOTA study phase 3

IOTA method	Variables or features
LR1 [10] (risks ⩾10% indicate malignancy)	(1) Personal history of ovarian cancer (yes, 1; no, 0), (2) current use of hormonal therapy (yes, 1; no, 0), (3) age of the patient (in years), (4) maximum diameter of lesion (in mm), (5) tender mass at examination (yes, 1; no, 0), (6) ascites (yes, 1; no, 0), (7) blood flow in papillary projection (yes, 1; no, 0), (8) purely solid tumour, (9) maximum diameter of the largest solid component (in mm, but with no increase >50 mm), (10) irregular internal cyst walls (yes, 1; no, 0), (11) acoustic shadows (yes, 1; no, 0), and (12) colour flow score (1–4, where 1 is no flow and 4 is maximum flow). The mathematical formula is presented in Supplementary Appendix.
LR2[10], (risks ⩾10% indicate malignancy)	(1) Ascites (yes, 1; no, 0), (2) blood flow in papillary projection (yes, 1; no, 0), (3) maximum diameter of the largest solid component (in mm, but with no increase >50 mm), (4) irregular internal cyst walls (yes, 1; no, 0), (5) acoustic shadows (yes, 1; no, 0), (6) age of the patient (in years). The mathematical formula is presented in Supplementary Appendix.
IOTA SRs[11,a]	Benign features: unilocular tumour (B1), largest diameter of largest solid component <7 mm (B2), acoustic shadows (B3), smooth multilocular tumour with largest diameter <100 mm (B4), no intratumoural blood flow at colour or power Doppler (B5). Malignant features: irregular solid tumour (M1), ascites (M2), At least four papillary projections (M3), irregular multilocular solid tumour with largest diameter ⩾100 mm (M4), very strong intratumoural blood flow at colour or power Doppler (M5).
IOTA SDs[12,b]	Benign descriptors: unilocular tumour with ground glass echogenicity in a premenopausal woman; unilocular tumour with mixed echogenicity and acoustic shadows in a premenopausal woman; unilocular anechoic tumour with regular walls and maximum diameter of lesion <10 cm; remaining unilocular tumours with regular walls. Malignant descriptor: tumour with ascites and at least moderate colour Doppler blood flow in a postmenopausal woman; age >50 years and CA-125 >100 U ml^{-1}.

Abbreviations: IOTA = international ovarian tumour analysis; LR1 = logistic regression model 1; LR2 = logistic regression model 2; SDs = simple descriptors; SR = simple rules.
[a]A mass is classified as malignant if at least one M-feature and none of the B-features are present and vice versa. If no B or M features are present, or if both B and M features are present, then the rules are considered inconclusive (unclassifiable mass), and a second stage test should be used in the unclassifiable tumours.
[b]A mass classified as malignant if at least one malignant descriptor and none of the benign descriptors are present and vice versa. If no benign or malignant descriptors are present, or if both benign and malignant descriptors are present, then the descriptors are inconclusive (unclassifiable mass), and a second stage test should be used in the unclassifiable tumours.

which SRs are inconclusive; SDs as a first stage test, LR2 for tumours in which the SDs are not applicable and SA for masses in which LR2 yields a predicted risk of ⩾5% but <25%.

Reference standard. The reference standard was the histologic classification of the excised mass as malignant or benign. Histological examination was carried out at the local centre. Central pathology review was not performed because in previous IOTA studies no significant differences in reported outcomes were observed between local and central pathology reports (Timmerman *et al*, 2005). Malignant tumours were classified according to the criteria recommended by the International Federation of Gynaecology and Obstetrics (Heintz *et al*, 2003). Borderline ovarian tumours were classified as malignant. The pathologist was blinded to the prediction outcomes of the index tests being compared.

Statistical analysis. We evaluated all strategies in terms of their ability to discriminate between benign and malignant masses. For the logistic regression models LR1 and LR2 (for details, see Table 1) and for RMI, the area under the receiver-operating characteristic curve (AUC) was computed. Using the six levels of diagnostic confidence, an AUC could also be constructed for SA. For all strategies, we calculated sensitivity, specificity, positive and negative likelihood ratio (LR+ and LR–) and diagnostic odds ratio (DOR; Deeks, 2001). To do this for LR1, LR2 and RMI, we used the cutoffs suggested in previous work (i.e., risk of malignancy ⩾10% indicating malignancy when using LR1 and LR2, and RMI >200 indicating malignancy). To recognise that performance may differ across centres, results were computed using meta-analysis techniques (Riley *et al*, 2008; Macaskill *et al*, 2010; Van Klaveren *et al*, 2014). To obtain the average AUC and DOR estimates, random effects meta-analysis was performed, using the logit of the AUC or the log of the DOR as the outcome variable. Sensitivity and specificity were modelled simultaneously using random centre effects. LR+ and LR– were computed based on the estimated average sensitivity and specificity levels. Forest plots for LR2, SRs and RMI were used to present centre-specific and combined results. Subgroup analyses for RMI and the most extensively validated IOTA methods (i.e., LR2 and SRs) (Timmerman *et al*, 2010a; Hartman *et al*, 2012; Nunes *et al*, 2012, 2013; Alcázar *et al*, 2013; Sayasneh *et al*, 2013a, b) were performed for pre- and postmenopausal women.

For LR2, we also assessed calibration, that is, we tested the extent to which the estimated risks of malignancy corresponded to the observed prevalence of malignancy. This was carried out by constructing parametric (logistic) calibration curves per centre (Cox, 1958; Steyerberg, 2009; Bouwmeester *et al*, 2013). Risk of Malignancy Index and SRs do not provide risk estimates but comparable centre-specific curves were obtained for RMI and SRs in the following manner. For RMI, analogous logistic curves were constructed to link RMI values (based on log (RMI + 1)) to observed risks. For SRs, the proportion of malignant masses was calculated for each classification level (benign, inconclusive and malignant).

CA-125 is not a mandatory variable in the IOTA studies and by consequence information on CA-125 was missing in 40% of the patients. It is most likely that missing values mainly arose when investigators did not consider CA-125 measurement necessary given the clinical situation and the ultrasound appearance of the mass. We used multiple imputation to handle the missing values (Sterne *et al*, 2009). We used all patients from phases 1, 1b, 2 and 3 for the imputation analysis. The method is described in more detail in a Supplementary appendix and elsewhere (Van Calster *et al*, 2011).

Calculations were performed using SAS 9.3 (SAS Institute, Cary, NC, USA). Forest plots were created in R (www.r-project.org) using the rmeta package.

When writing this article, we used the Standards for Reporting of Diagnostic Accuracy (STARD) guidelines (Bossuyt *et al*, 2003).

RESULTS

In total, 2541 women with adnexal masses were enrolled in this study. One hundred thirty-eight women were excluded from the final data set. Reasons for exclusion were: an interval of >120 days

between ultrasonography and surgery ($n=66$), pregnancy ($n=31$), data errors that could not be solved by contacting the respective principal investigators ($n=28$) and incomplete final histology ($n=13$). The final data set included 2403 patients with 1423 (59%) benign and 980 (41%) malignant adnexal masses. There were 1049 postmenopausal patients (44%) and 1354 (56%) premenopausal patients. The prevalence of malignancy was 28% (378 out of 1354) in premenopausal patients and 57% (602 out of 1049) in postmenopausal patients.

The types of benign and malignant tumours based on histology and FIGO staging in the final data set are presented in Table 2. The most common benign diagnoses were endometrioma, serous cystadenoma and teratoma. There were 633 out of 2403 (26.3%) primary invasive ovarian cancers, 153 out of 2403 (6.4%) borderline ovarian tumours, 126 out of 2403 (5.2%) metastatic cancers in the ovaries and 68 out of 2403 (2.8%) rare primary invasive ovarian malignancies (e.g., granulosa cell tumour, Sertoli–Leydig cell tumour or dysgerminoma). Descriptive statistics for the variables included in LR1, LR2 and the SRs in benign and malignant adnexal masses are shown in Table 3.

The test performance of the IOTA diagnostic strategies, SA and RMI when using a meta-analysis approach on centre-specific data are presented in Table 4. The logistic regression models LR1 (AUC 0.930; 0.917–0.942) and LR2 (AUC 0.918; 0.905–0.930) had diagnostic performance similar to expert SA (AUC 0.914; 0.886–0.936). The AUC of RMI was 0.875 (0.853–0.894). The IOTA risk models LR1 and LR2, and strategies using various combinations of SRs, SDs and LR2 achieved sensitivity 90–96% and specificity 74–79% (Table 4). When expert SA was used in case SRs or SDs or both yielded an inconclusive result, or in the event that LR2 gave a risk $\geqslant 5\%$ but $<25\%$, specificity increased from 74–79% to 85–89% with slightly reduced sensitivity in most instances (Table 4). The sensitivity of RMI was 67% and the specificity 91%. LR2 yielded a risk of $\geqslant 5\%$ but $<25\%$ in 419 of the 2403 patients (17%), and 108 (26%) of these women had a malignant adnexal mass.

The IOTA SRs were applicable in 1846 patients (76.8%) and could be applied slightly more frequently in premenopausal (1055 out of 1354) (77.9%) than postmenopausal women (791 out of 1049) (75.4%). In total, 1090 tumours were classified as benign by the SRs, and this was correct in 1044 cases (95.8%), 756 tumours were classified as malignant by the SRs and this was correct in 674 cases (89.2%). When the SRs were applicable they achieved a sensitivity of 94% (674 out of 720) and a specificity of 93% (1044 out of 1126). The malignancy rate among tumours where the SRs yielded an inconclusive result was 46.7% (260 out of 557). A strategy that used SRs as a first stage test and classified all inconclusive cases as malignant yielded a sensitivity of 95% (95% CI 93–97%) and a specificity of 74% (95% CI 68–80%; Table 4). Using SA by an expert examiner when SRs yielded an inconclusive result lowered the sensitivity from 95 to 92% (89–94%) but increased the specificity from 74 to 89% (85–92%; Table 4).

The IOTA SDs could be applied in 1014 (42.2%) masses. The SDs classified 549 tumours (23% of all tumours in the study) as

Table 2. Overview of tumour types

Pathology	Frequency (*n*)	Percent (%)
Benign masses		
Endometrioma	344	14.3
Serous cystadenoma	259	10.8
Teratoma	231	9.6
Mucinous cystadenoma	183	7.6
Fibroma	130	5.4
Simple cyst or parasalpingeal cyst	106	4.4
Rare benign[a]	48	2.0
Hydrosalpinx or salpingitis	47	2.0
Functional cyst	40	1.7
Peritoneal pseudocyst	18	0.8
Abscess	17	0.7
Malignant masses		
Primary invasive stage I	128	5.3
Primary invasive stage II	47	2.0
Primary invasive stage III	397	16.5
Primary invasive stage IV	61	2.5
Borderline stage I	135	5.6
Borderline stage II	6	0.3
Borderline stage III	12	0.5
Rare primary invasive[b]	68	2.8
Metastatic	126	5.2
Total	2403	100

[a]For example, Brenner tumour or struma ovarii.
[b]For example, dysgerminoma, granulosa cell tumour, yolk sac tumour or malignant teratoma.

Table 3. Results with regard to the variables included in LR1, LR2 and the Simple Rules in benign and malignant adnexal masses

Variable	Statistics	Benign (1423, 59%)	Malignant (980, 41%)
Variables in LR1 and LR2			
Age (years)	Median (IQR)	44 (33–56)	57 (46–66)
Largest diameter of lesion (mm)	Median (IQR)	64 (47–90)	86 (56–126)
Solid components	*N* (%)	472 (33%)	915 (93%)
Largest diameter of solid component if present (mm)	Median (IQR)	28 (13–54)	59 (37–87)
Colour score (1–4)			
Colour score 1	*N* (%)	574 (40)	32 (3)
Colour score 2	*N* (%)	563 (40)	199 (20)
Colour score 3	*N* (%)	239 (17)	442 (45)
Colour score 4	*N* (%)	47 (3)	307 (31)
Ascites	*N* (%)	18 (1%)	322 (33%)
Papillations with detectable blood flow	*N* (%)	55 (4%)	160 (16%)
Irregular cyst walls	*N* (%)	385 (27%)	572 (58%)
Acoustic shadows	*N* (%)	265 (19%)	34 (3%)
Tender mass at ultrasound examination	*N* (%)	233 (16%)	111 (11%)
Current use of hormonal therapy	*N* (%)	153 (11%)	54 (6%)
Personal history of ovarian cancer	*N* (%)	14 (1%)	30 (3%)
Solid tumour	*N* (%)	154 (11%)	473 (48%)
Variables in the simple rules			
Benign ultrasound features in the simple rules			
Unilocular tumour (B1)	*N* (%)	595 (42%)	5 (0.5%)
Largest diameter of largest solid component <7 mm (B2)	*N* (%)	40 (3%)	2 (0.2%)
Acoustic shadows (B3)	*N* (%)	265 (19%)	34 (3%)
Smooth multilocular tumour with largest diameter <100 mm (B4)	*N* (%)	224 (16%)	13 (1%)
No intratumoural blood flow at colour or power Doppler (B5)	*N* (%)	574 (40%)	32 (3%)
Malignant ultrasound features in the simple rules			
Irregular solid tumour (M1)	*N* (%)	16 (1%)	189 (19%)
Ascites (M2)	*N* (%)	18 (1%)	322 (33%)
At least 4 papillary projections (M3)	*N* (%)	27 (2%)	91 (9%)
Irregular multilocular solid tumour with largest diameter $\geqslant 100$ mm (M4)	*N* (%)	40 (3%)	153 (16%)
Very strong intratumoural blood flow at colour or power Doppler (M5)	*N* (%)	47 (3%)	307 (31%)

Abbreviations: IQR = interquartile range; LR1 = logistic regression model 1; LR2 = logistic regression model 2.

Table 4. Test performance of the IOTA diagnostic strategies, subjective assessment and RMI when using a meta-analysis approach on centre-specific data

Approach	AUC (95% CI)	Sens, % (95% CI)	Spec, % (95% CI)	LR+	LR–	DOR (95% CI)
One-step strategies						
LR1	0.930 (0.917–0.942)	93.7 (91.4–95.4)	77.6 (70.9–83.0)	4.17	0.08	40.8 (30.0–55.4)
LR2	0.918 (0.905–0.930)	90.2 (86.9–92.8)	78.9 (73.2–83.7)	4.28	0.12	31.2 (23.1–42.2)
SA	0.914 (0.886–0.936)	92.5 (89.4–94.8)	87.7 (83.2–91.2)	7.53	0.09	72.9 (49.8–107)
RMI	0.875 (0.853, 0.894)	67.1 (61.4–72.4)	90.6 (87.3–93.1)	7.15	0.36	17.5 (13.1–23.4)
SRMal	NA	95.3 (93.1–96.9)	74.1 (67.7–79.7)	3.68	0.06	49.1 (34.9–69.0)
Two-step strategies						
SR + SA	NA	91.8 (89.1–93.9)	89.0 (85.2–92.0)	8.38	0.09	75.7 (55.6–103)
LR2 + SA	NA	92.3 (89.5–94.5)	84.8 (80.4–88.3)	6.06	0.09	58.7 (43.4–79.4)
SD + SRMal	NA	95.7 (93.5–97.1)	73.6 (66.7–79.5)	3.62	0.06	50.5 (35.7–71.6)
SD + LR2	NA	91.1 (88.1–93.5)	78.1 (72.4–82.9)	4.17	0.11	32.8 (24.6–43.7)
SD + SA	NA	93.0 (90.0–95.1)	86.5 (81.8–90.1)	6.88	0.08	68.5 (47.7–98.3)
Three-step strategies						
SD + SR + SA	NA	92.5 (89.6–94.6)	87.6 (83.5–90.7)	7.44	0.09	70.7 (51.7–96.5)
SD + LR2 + SA	NA	93.1 (90.5–95.0)	83.7 (79.2–87.4)	5.71	0.08	57.6 (42.3–78.6)

Abbreviations: AUC = area under the receiver-operating characteristics curve; CI = confidence interval; DOR = diagnostic odds ratio; IOTA = international ovarian tumour analysis; LR+ = positive likelihood ratio; LR– = negative likelihood ratio; LR1 = logistic regression model 1; LR2 = logistic regression model 2; LR2 + SA = LR2 as a first stage test and SA for tumours in which LR2 yields a predicted risk of malignancy of ⩾5% but <25%; NA = not applicable; RMI = risk of malignancy index; RMI-1 = risk of malignancy index-1; SA = subjective assessment; SD = simple descriptor; SD + SRMal = SDs as a first stage test, SRs for tumours unclassifiable by the descriptors with all tumours in which SRs are inconclusive being classified as malignant; SD + LR2 = SDs as a first stage test and LR2 for those tumours in which the descriptors are not applicable; SD + SA = SDs as a first stage test and SA for those tumours in which the SDs are not applicable; SD + SR + SA = SDs as a first stage test, SRs for tumours in which the descriptors are not applicable, and SA for masses in which the SRs are inconclusive; SD + LR2 + SA = SDs as a first stage test, LR2 for tumours in which the descriptors are not applicable, and SA for masses in which LR2 yields a predicted risk of ⩾5% but <25%; Sens = sensitivity; Spec = specificity; SR = simple rules; SRMal = SRs as a first stage test with all tumours in which SRs are inconclusive being classified as malignant; SR + SA = SRs as a first stage test and SA for tumours in which the SRs are inconclusive.

benign of which 2 (0.4%) turned out to be malignant. The two misclassified malignancies were stage I borderline tumours. The SDs classified 465 tumours (19% of all tumours in the study) as malignant, and 430 (92.5%) proved to be so. The 35 benign tumours misclassified as malignant by the SDs consisted of 11 serous cystadenomas, 10 fibromas, 7 mucinous cystadenomas, 4 rare benign tumours, 2 teratomas and 1 functional cyst. A total of 1389 (58%) tumours could not be categorised with the SDs. The SRs could be applied in 66% (912 out of 1389) of the tumours unclassifiable by the descriptors. The combination of SDs with SRs characterised 80% (1926 out of 2403) of all masses as benign or malignant. When a three-step strategy was applied (SDs as first stage test, SRs in tumours unclassifiable by the descriptors and SA for masses in which the SRs were inconclusive), sensitivity and specificity were 93% (95% CI 90–95%) and 88% (95% CI 84–91%), respectively (Table 4).

Table 5 shows the test performance of LR2, SRs and RMI in pre- and postmenopausal patients when using a meta-analysis approach of centre-specific data. In both pre- and postmenopausal patients, the IOTA strategies had higher sensitivity and lower specificity than RMI. The use of SA for masses not classifiable by the SRs appeared to resolve the differences in specificity.

The sensitivity and specificity of LR2, SRs and RMI for histological subtypes of malignant disease and the absolute number of false-negative results for histological subtypes of malignancy are presented in Supplementary Tables S1–S4. The sensitivity with regard to borderline tumours, FIGO stage I invasive cancer and metastatic disease was much higher for the main IOTA approaches than for RMI, and the AUCs for LR2 were larger than those for RMI for these subtypes of malignancy.

Figure 1 and Supplementary Figure S1 illustrate the variation in number of included masses, prevalence of malignancy and inter-centre differences in test performance (sensitivity and specificity) for LR2, SRs combined with subjective expert assessment, SRs and classifying inconclusive tumours as malignant, and RMI. The malignancy rate varied between 0 and 69%, whereas the number of enrolled cases per centre ranged from 6 to 443. For LR2 and SRs, differences between centres in sensitivity were smaller than differences in specificity, whereas the inverse held true for RMI. Both IOTA methods had a higher sensitivity for cancer than RMI, irrespective of the prevalence of malignancy. Discrimination (AUCs) for LR2 was consistent in both oncology and non-oncology centres with a few exceptions for centres that enrolled a

Table 5. Test performance of LR2, Simple Rules and RMI in pre- and postmenopausal patients using a meta-analysis approach on centre-specific data

Diagnostic method	AUC (95% CI)	Sens, % (95% CI)	Spec, % (95% CI)
Premenopausal patients			
LR2	0.908 (0.886–0.926)	85 (78–90)	82 (77–87)
SRMal		95 (91–97)	77 (70–83)
SR + SA		92 (86–95)	91 (87–94)
RMI	0.867 (0.837–0.892)	53 (45–61)	94 (92–96)
Postmenopausal patients			
LR2	0.897 (0.872–0.917)	94 (92–96)	65 (58–71)
SRMal		96 (93–97)	66 (59–73)
SR + SA		93 (90–95)	83 (78–87)
RMI	0.850 (0.805–0.887)	78 (72–83)	81 (76–85)

Abbreviations: AUC = area under the receiver-operating characteristics curve; CI = confidence interval; IOTA = international ovarian tumour analysis; LR2 = logistic regression model 2; RMI = Risk of Malignancy Index; RMI-1 = risk of malignancy index-1; Sens = sensitivity; Spec = specificity; SRMal = a one-step strategy using the IOTA Simple Rules as a first stage test and classifying tumours where the simple rules yield an inconclusive result as malignant; SR + SA = a two-stage strategy using the IOTA simple rules as a first stage test and using subjective assessment for those tumours where the simple rules yield an inconclusive result.

very small number of cases (Supplementary Figure S2). Discrimination for RMI showed some variation between the centres (Supplementary Figure S3). The summary estimates of test performance of LR2, SRs and RMI were similar irrespective of whether it was estimated using pooled data or meta-analysis. However, pooling underestimates uncertainty, while uncertainty is appropriately addressed by adopting meta-analysis techniques.

Figure 2 shows the results of calibration for LR2, RMI and SRs for the nine centres that contributed the largest number of patients. Calibration results differed between centres. This means that for a specific prediction from a diagnostic test (LR2 risk, RMI value or SRs category) the observed prevalence of malignancy varies between centres. For LR2, the risk of malignancy was underestimated in seven of the nine centres (calibration curves above the diagonal), slightly overestimated in one centre and perfectly calibrated in one centre. The prevalence of malignancy in women with a RMI score of 200 varied between centres from 30 to 70%, and RMI values of 200 or more were associated with high malignancy rates.

DISCUSSION

This comparison of IOTA risk prediction models and diagnostic strategies in different clinical environments using a meta-analysis approach showed excellent test performance for all IOTA methods to characterise adnexal masses before surgery. All IOTA strategies manifested better discrimination than RMI. In addition, we have demonstrated inter-centre differences in test performance and calibration for LR2, SRs and RMI. Our use of meta-analysis techniques to summarise data did not meaningfully change the summary measures of performance from those obtained with a standard pooled analysis but gave wider confidence intervals properly reflecting the uncertainty caused by differences between centres.

The strengths of this report include the use of a rigorous prospective ultrasound protocol with agreed terms, measurement techniques and definitions; the use of advanced statistical methods to synthesise multicentre data and report summary estimates of test accuracy and calibration, thereby minimising the risk that results are overly influenced by a single centre recruiting many more patients than others; and the large number of patients, the many participating centres and the different types of participating centres making our results highly likely to be generalisable. A limitation of our study is that the SRs, SDs and the two-step and three-step strategies were not directly applied when scanning the patients. Instead, >40 clinical and ultrasound variables were prospectively collected from each patient and later incorporated in the SRs or SDs, or synthesised to become descriptors in the SDs or features in the SRs. Although this may not have influenced the performance of SA or of LR1 or LR2, it could have affected that of the other tests and of the two-step and three-step strategies. A second limitation is that information on CA-125 was missing in 40% of cases. We solved this by using multiple imputation (Sterne *et al*, 2009; Van Calster *et al*, 2011). Two sensitivity analyses using only complete cases for CA-125 confirmed the difference in test performance in favour of the IOTA methods (LR2 and SRs) (Supplementary Tables S5 and S6). However, these approaches are biased because CA-125 is more often missing in tumours that are easy to diagnose and more likely to be benign (Supplementary Table S7). This explains why model performance was slightly poorer for all methods when they were tested only in cases with available CA-125 results. For this reason, multiple imputation is generally considered a more appropriate method to deal with missingness than to analyse only data with complete information (Sterne *et al*, 2009). A third limitation is that most of the patients in the IOTA phase 3 study were scanned by the same experienced examiners as in the centres where the IOTA methods were developed, or by examiners that had already adopted the IOTA examination technique and terminology. This may explain why the results of IOTA phase 3 confirm those of previous IOTA studies that showed excellent test performance of all IOTA strategies (Timmerman *et al*, 2005, 2007, 2008, 2010a, b; Van Holsbeke *et al*, 2009, 2012; Ameye *et al*, 2012; Kaijser *et al*, 2013; Sayasneh *et al*, 2013a, b). On the other hand, validation studies of LR1, LR2 and SRs performed outside IOTA studies reported similar results (Hartman *et al*, 2012; Nunes *et al*, 2012, 2013; Alcázar *et al*, 2013), and there is now evidence that the IOTA strategies retain their performance in the hands of sonographers and relatively inexperienced doctors (Hartman *et al*, 2012; Nunes *et al*, 2012, 2013; Alcázar *et al*, 2013; Sayasneh *et al*, 2013a, b).

Our results showing that the IOTA methods and strategies have excellent ability to discriminate between benign and malignant adnexal masses and are superior to RMI in this regard are in line with other validation studies (Timmerman *et al*, 2010a; Van Holsbeke *et al*, 2012; Sayasneh *et al*, 2013a, b) and a recent systematic review (Kaijser *et al*, 2014). The conclusion of the review was that an evidence-based approach to the preoperative characterisation of adnexal masses should incorporate the use of IOTA SRs or LR2 instead of RMI, particularly in women of reproductive age (Kaijser *et al*, 2014).

Our multicentre study demonstrated differences in test performance between centres for LR2, SRs and RMI. However, in all centres, also in those with a low observed prevalence of malignancy, the sensitivity with regard to malignancy was much higher for the IOTA methods than for RMI. The overall discriminative capacity (AUC) for LR2 and RMI did not seem to be affected by cancer prevalence. However, our study highlighted important differences in calibration results for LR2, SRs and RMI. Type of centre appeared to contribute to these differences: oncology centres have a higher prevalence of malignant tumours and suffered from underestimation of the predicted risk. This was also noticeable for SRs and RMI although these methods do not directly provide an estimated risk. For example, for RMI we can derive that the implicit average risk at an RMI value of 200 is 54%. This implies that at this (commonly used) cutoff, in some centres patients with a risk of malignancy of $>50\%$ may be classified as low risk.

We did not undertake a full meta-regression analysis (van Houwelingen *et al*, 2002) to explain in detail the inter-centre differences in results for test performance and calibration as this is beyond the scope of this article. The most plausible explanations are differences in study populations (e.g., patients' age, body mass index or tumour mix), equipment and examiners' use of the IOTA terms. However, the variation between centres of the observed (true risk) versus predicted risks for ovarian cancer revealed by this meta-analysis highlights that caution is needed when interpreting and using diagnostic test results (that is, risks) for individual patient management within the context of a single centre. Future studies should explore the reasons for differences in diagnostic performance and calibration of diagnostic approaches between different centres, the final aim being to improve risk prediction for ovarian cancer.

Subjective assessment of grey scale and Doppler ultrasound findings by a very experienced ultrasound examiner has been suggested to be the preferred approach to characterise adnexal masses (Valentin *et al*, 2001). Unfortunately, most gynaecologists, radiologists and sonographers have limited experience with the use of SA of ultrasound images to discriminate between benign and malignant adnexal masses. As SRs, SDs and LR2 have been shown to perform very well in the hands of both sonographers and gynaecologists with limited ultrasound experience (Hartman *et al*, 2012; Nunes *et al*, 2012, 2013; Alcázar *et al*, 2013; Sayasneh *et al*, 2013a, b), they could be used as first stage tests, and patients with

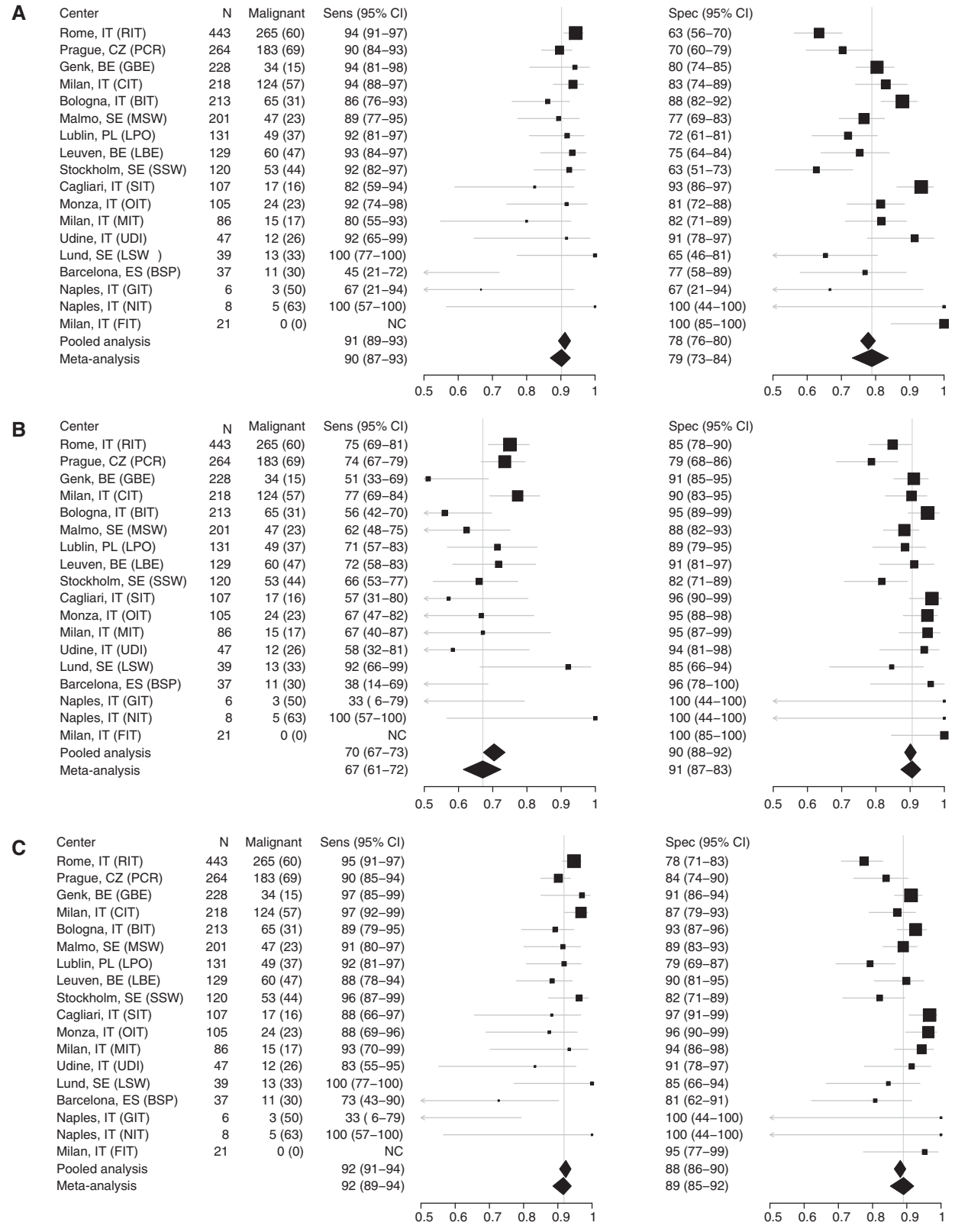

Figure 1. The sensitivity (Sens) and specificity (Spec) for LR2 (A), Risk of Malignancy Index (B) and a two-stage strategy using SRs as a first stage test and using SA for tumours in which the SRs are inconclusive (C) per contributing centre and for all centres combined using a meta-analysis approach and pooled data. NC = not computable. Numbers in brackets denote the prevalence (%) of malignant masses in each centre. Oncology centres were: University Hospitals Leuven, Belgium (LBE); Universita Cattolica del Sacro Cuore, Rome, Italy (RIT); Ospedale San Gerardo, Monza, Italy (OIT); General Faculty Hospital, Prague, Czech Republic (PCR); Istituto Europeo di Oncologia, Milan, Italy (CIT); Medical University Lublin, Poland (LPO); Karolinska University Hospital, Stockholm, Sweden (SSW); Skåne University Hospital Lund, Sweden (LSW); Universita degli Studi di Udine, Italy (UDI); Istituto Nazionale dei Tumori, Naples, Italy (GIT); University of Bologna, Italy (BIT). Non-oncology centres were: Skåne University Hospital Malmö, Sweden (MSW), Ziekenhuis Oost-Limburg, Genk, Belgium (GBE); Ospedale San Giovanni di Dio, Cagliari, Italy (SIT); DCS Sacco University of Milan, Italy (MIT); Universita degli Studi di Napoli, Naples, Italy (NIT); Institut Universitari Dexeus, Barcelona, Spain (BSP); Ospedale dei Bambini Vittore Buzzi, Milan, Italy (FIT).

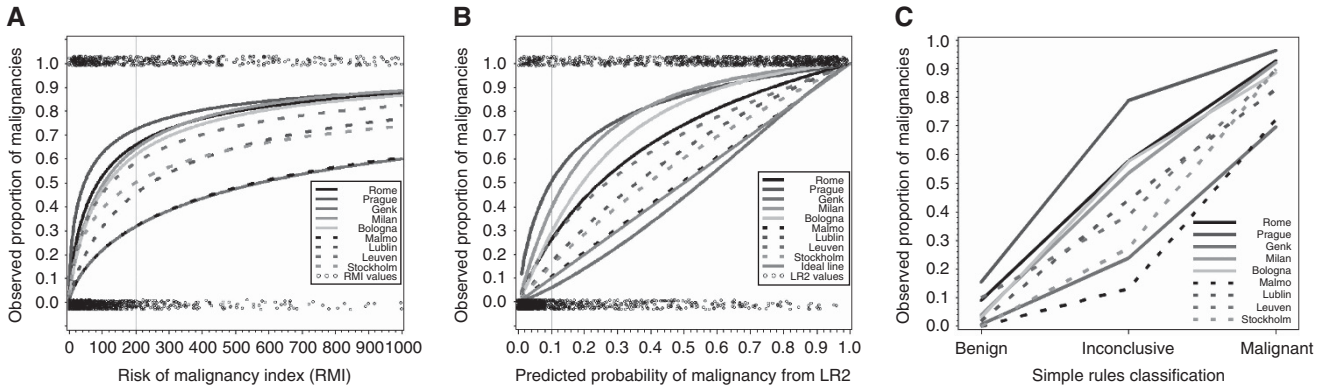

Figure 2. Centre-specific calibration curves (Cox logistic recalibration) for RMI (**A**), LR2 (**B**), and proportions for SRs (**C**). Vertical lines crossing the x axis for RMI (200) and LR2 (0.1, i.e., 10% risk) represent the original cutoff to define malignant disease. Scatter plots above and below the calibration curves for RMI and LR2 represent the distribution of predicted risks for LR2 and values for RMI for benign and malignant tumours, respectively. The x axis for RMI is limited to 1000. Oncology centres were: Rome, Prague, Milan, Leuven, Stockholm, Bologna and Lublin. Non-oncology centres were: Malmö and Genk.

inconclusive or equivocal results of the first stage test could be referred for SA by an experienced ultrasound examiner.

Each IOTA strategy has its own advantages and disadvantages. For example, LR1 and LR2 give a continuous result (a risk estimate) for which the cut off to diagnose malignancy can be varied depending on the context. The SRs are easier to apply than LR1 and LR2, which require a computer or mobile application, but do not offer the flexibility of LR1 and LR2. However, all of these approaches can be used to either classify all patients or classify a majority of patients while referring a subset of patients for further testing. Using SDs as a first stage test offers no substantial advantages in test performance over the other IOTA strategies that we evaluated in this work. However, referring patients to expert examiners with masses in which the SRs do not apply or with equivocal results of LR2 is advantageous as it leads to a reduction in the false-positive rate, while only minimally decreasing the sensitivity. In this study as well as in IOTA phase 2 data, SRs were inconclusive in 23% of patients, whereas LR2 results were equivocal in 17–18% of the same patients (Timmerman *et al*, 2010a; Van Calster *et al*, 2012). In other validation studies of the SRs, fewer patients had inconclusive results with reported percentages between 11 and 21% in different populations (Fathallah *et al*, 2011; Hartman *et al*, 2012; Alcázar *et al*, 2013; Sayasneh *et al*, 2013b).

The results of IOTA3 show that IOTA methods result in better discrimination of adnexal pathology before surgical treatment irrespective of the prevalence of malignant disease. Therefore, the application of IOTA risk models or rules provide a rational basis for referral of patients with a mass classified as malignant to specialist oncology services.

ACKNOWLEDGEMENTS

We are grateful to all study centres, investigators and patients for their collaboration. BVC is a postdoctoral fellow of the Research Foundation—Flanders. LW is supported by a PhD grant of the Flanders' Agency for Innovation by Science and Technology (IWT Vlaanderen). This study is supported by FWO (project G049312N) and IWT (IWT-TBM 070706-IOTA3, the Swedish Medical Research Council (grants no. K2006-73X-11605-11-3); funds administered by Skåne University Hospital; Allmänna Sjukhusets i Malmö Stiftelse för bekämpande av cancer (the Malmö General Hospital Foundation for fighting against cancer); Landstingsfinansierad regional forskning and ALF-medel (i.e., two Swedish governmental grants from the region of Scania); Funds administered by Skåne University Hospital. TB is supported by Imperial Healthcare NHS Trust NIHR Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

DT, AT, BVC, IV, TB and LV conceived of the study and its design. AT, DFi, CVH, DFr, LS, EE, AC, SG, RF, FPGL, LV, JK and DT enrolled patients and acquired data. BVC, LW, CVH, JK and DT were involved in data cleaning. BVC analysed the data, with support from LW. BVC, LW, LV, AT, JK, TB and DT were involved in data interpretation. AT, JK, LW, BVC, LV, TB and DT wrote the first draft of the manuscript, which was then critically reviewed and revised by all other co-authors. All authors approved the final version of the manuscript for submission, and agree to be accountable for all aspects of the work relating to accuracy and integrity.

REFERENCES

Alcázar JL, Pascual MA, Olartecoechea B, Graupera B, Aubá M, Ajossa S, Hereter L, Julve R, Gastón B, Peddes C, Sedda F, Piras A, Saba L, Guerriero S (2013) IOTA simple rules for discriminating between benign and malignant adnexal masses: prospective external validation. *Ultrasound Obstet Gynecol* **42**(4): 467–471.

Ameye L, Timmerman D, Valentin L, Paladini D, Zhang J, Van Holsbeke C, Lissoni AA, Savelli L, Veldman J, Testa AC, Amant F, Van Huffel S, Bourne T (2012) Clinically oriented three-step strategy for assessment of adnexal pathology. *Ultrasound Obstet Gynecol* **40**(5): 582–591.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis PP, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, De Vet HC (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Ann Intern Med* **138**: 40–44.

Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y (2013) Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med Res Methodol* **13**: 19.

Cox DR (1958) Two further applications of a model for binary regression. *Biometrika* **45**: 562–565.

Deeks JJ (2001) Systematic reviews in health care: systematic reviews of evaluations of diagnostic and screening tests. *BMJ* **323**(7305): 157–162.

Education, Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (2006) Minimum training recommendations for the practice of medical ultrasound. *Ultraschall in der Medizin* **27**(1): 79–105.

Fathallah K, Huchon C, Bats AS, Metzger U, Lefrere-Belda MA, Bensaid C, Lecuru F (2011) External validation of simple ultrasound rules of Timmerman on 122 ovarian tumors. *Gynecologie Obstetrique Fertilite* **39**(9): 477–481.

Hartman CA, Juliato CR, Sarian LO, Toledo MC, Jales RM, Morais SS, Pitta DD, Marussi EF, Derchain S (2012) Ultrasound criteria and CA 125 as predictive variables of ovarian cancer in women with adnexal tumors. *Ultrasound Obstet Gynecol* **40**(3): 360–366.

Heintz APM, Odicino F, Maisonneuve P, Beller U, Benedet JL, Creasman WT, Ngan HY, Pecorelli S (2003) Carcinoma of the ovary: 25th annual report on the results of treatment in gynecological cancer. *Int J Gynaecol Obstet* **83**(suppl 1): S135–S137.

Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG (1990) A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. *Br J Obstet Gynaecol* **97**(10): 922–929.

Kaijser J, Sayasneh A, Van Hoorde K, Ghaem-Maghami S, Bourne T, Timmerman D, Van Calster B (2014) Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. *Human Reprod Update* **20**(3): 449–462.

Kaijser J, Van Gorp T, Van Hoorde K, Van Holsbeke C, Sayasneh A, Vergote I, Bourne T, Timmerman D, Van Calster B (2013) A comparison between an ultrasound based prediction model (LR2) and the Risk of Ovarian Malignancy Algorithm (ROMA) to assess the risk of malignancy in women with an adnexal mass. *Gynecol Oncol* **129**(2): 377–383.

Macaskill P, Gatsonis C, Deeks JJ, Harbord R, Takwoingi Y (2010) Chapter 10: analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C (eds) *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration*. Available from http://srdta.cochrane.org/.

Miller RW, Ueland FR (2012) Risk of malignancy in sonographically confirmed ovarian tumors. *Clin Obstet Gynecol* **55**(1): 52–64.

Moore RG, McMeekin DS, Brown AK, DiSilvestro P, Miller MC, Allard WJ, Gajewski W, Kurman R, Bast Jr RC, Skates SJ (2009) A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. *Gynecol Oncol* **112**: 40–46.

Nunes N, Ambler G, Hoo WL, Naftalin J, Foo X, Widschwendter M, Jurkovic D (2013) A Prospective validation of the IOTA logistic regression models (LR1 and LR2) in comparison to subjective pattern recognition for the diagnosis of ovarian cancer. *Int J Gynecol Cancer* **23**(9): 1583–1589.

Nunes N, Yazbek J, Ambler G, Hoo W, Naftalin J, Jurkovic D (2012) A prospective evaluation of the IOTA logistic regression model (LR2) for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol* **40**(3): 355–359.

Riley RD, Dodd SR, Craig JV, Thompson JR, Williamson PR (2008) Meta-analysis of diagnostic test studies using individual patient data and aggregate data. *Stat Med* **27**(29): 6111–6136.

Sayasneh A, Kaijser J, Preisler J, Johnson S, Stalder C, Husicka R, Guha S, Naji O, Abdallah Y, Raslan F, Drought A, Smith AA, Fotopoulou C, Ghaem-Maghami S, Van Calster B, Timmerman D, Bourne T (2013a) A multicenter prospective external validation of the diagnostic performance of IOTA simple descriptors and rules to characterize ovarian masses. *Gynecol Oncol* **130**(1): 140–146.

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 (2013b) Multicenter external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer* **108**(12): 2448–2454.

Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338**: b2393.

Steyerberg EW (2009) *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health)*. Springer: New York.

Timmerman D, Ameye L, Fischerova D, Epstein E, Melis GB, Guerriero S, Van Holsbeke C, Savelli L, Fruscio R, Lissoni AA, Testa AC, Veldman J, Vergote I, Van Huffel S, Bourne T, Valentin L (2010a) Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. *BMJ* **341**: c6839.

Timmerman D, Testa AC, Bourne T, Ameye L, Jurkovic D, Van Holsbeke C, Paladini D, Van Calster B, Vergote I, Van Huffel S, Valentin L (2008) Simple ultrasound-based rules for the diagnosis of ovarian cancer. *Ultrasound Obstet Gynecol* **31**(6): 681–690.

Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L. International Ovarian Tumor Analysis Group (2005) 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* **23**: 8794–8801.

Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. International Ovarian Tumor Analysis (IOTA) Group (2000) 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* **16**: 500–505.

Timmerman D, Van Calster B, Jurkovic D, Valentin L, Testa AC, Bernard JP, Van Holsbeke C, Van Huffel S, Vergote I, Bourne T (2007) Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors. *J Clin Oncol* **25**: 4194–4200.

Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, Van Holsbeke C, Fruscio R, Czekierdowski A, Jurkovic D, Savelli L, Vergote I, Bourne T, Van Huffel S, Valentin L (2010b) 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* **36**(2): 226–234.

Valentin L, Hagen B, Tingulstad S, Eik-Nes S (2001) Comparison of 'pattern recognition' and logistic regression models for discrimination between benign and malignant pelvic masses: a prospective cross-validation. *Ultrasound Obstet Gynecol* **18**: 357–365.

Van Calster B, Timmerman D, Valentin L, McIndoe A, Ghaem-Maghami S, Testa AC, Vergote I, Bourne T (2012) Triaging women with ovarian masses for surgery: observational diagnostic study to compare RCOG guidelines with an International Ovarian Tumour Analysis (IOTA) group protocol. *BJOG* **119**(6): 662–671.

Van Calster B, Valentin L, Van Holsbeke C, Zhang J, Jurkovic D, Lissoni AA, Testa AC, Czekierdowski A, Fischerová D, Domali E, Van de Putte G, Vergote I, Van Huffel S, Bourne T, Timmerman D (2011) A novel approach to predict the likelihood of specific ovarian tumor pathology based on serum CA-125: a multicenter observational study. *Cancer Epidemiol Biomarkers Prev* **20**(11): 2420–2428.

Van Holsbeke C, Van Calster B, Bourne T, Ajossa S, Testa AC, Guerriero S, Fruscio R, Lissoni AA, Czekierdowski A, Savelli L, Van Huffel S, Valentin L, Timmerman D (2012) External validation of diagnostic models to estimate the risk of malignancy in adnexal masses. *Clin Cancer Res* **18**(3): 815–825.

Van Holsbeke C, Van Calster B, Testa AC, Domali E, Lu C, Van Huffel S, Valentin L, Timmerman D (2009) Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. *Clin Cancer Res* **15**: 684–691.

van Houwelingen HC, Arends LR, Stijnen T (2002) Advanced methods in meta-analysis: multivariate approach and meta-regression. *Stat Med* **21**(4): 589–624.

Van Klaveren D, Steyerberg EW, Perel P, Vergouwe Y (2014) Assessing discriminative ability of risk models in clustered data. *BMC Med Res Methodol* **14**(1): 5.

Verleye L, Vergote I, van der Zee AG (2010) Patterns of care in surgery for ovarian cancer in Europe. *Eur J Surg Oncol* **36**(Suppl 1): S108–S114.

Woo YL, Kyrgiou M, Bryant A, Everett T, Dickinson HO (2012) Centralisation of services for gynaecological cancers—a Cochrane systematic review. *Gynecol Oncol* **126**(2): 286–290.

Supplementary Information accompanies this paper on British Journal of Cancer website (http://www.nature.com/bjc)