FULL PAPER

BJC

British Journal of Cancer (2013) 108, 2448–2454 | doi: 10.1038/bjc.2013.224

Keywords: Biomarkers; CA125 antigen; decision support techniques; ovarian cancer; ovarian neoplasm; ultrasonography

Multicentre external validation of IOTA prediction models and RMI by operators with varied training

A Sayasneh*[,1,2], L Wynants[3,4], J Preisler[2], J Kaijser[5], S Johnson[6], C Stalder[2], R Husicka[2], Y Abdallah[2], F Raslan[7], A Drought[7], A A Smith[8], S Ghaem-Maghami[1], E Epstein[9], B Van Calster[10], D Timmerman[3,10] and T Bourne[1,2,10]

[1]Department of Cancer and Surgery, Imperial College London, Hammersmith Campus, Du Cane Road, London W12 0HS, UK; [2]Early Pregnancy and Acute Gynecology Unit, Queen Charlottes and Chelsea Hospital, Imperial College London, Du Cane Road, London W12 0HS, UK; [3]Department of Electrical Engineering-ESAT, SCD-SISTA, KU Leuven, B-3000 Leuven, Belgium; [4]iMinds Future Health Department, KU Leuven, B-3000 Leuven, Belgium; [5]Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Herestraat 49, B-3000 Leuven, Belgium; [6]Southampton University Hospitals, Princess Anne Hospital, Coxford Road, Southampton SO16 6YD, UK; [7]West Middlesex University Hospital, Twickenham Road, Isleworth, Middlesex TW7 6AF, UK; [8]Department of Ultrasound, Queen Charlotte's and Chelsea Hospital, Du Cane Road, London W12 0HS, UK; [9]Department of Obstetrics and Gynecology, Karolinska University Hospital, S-171 76 Stockholm, Sweden and [10]Department of Development and Regeneration, KU Leuven, B-3000 Leuven, Belgium

Background: Correct characterisation of ovarian tumours is critical to optimise patient care. The purpose of this study is to evaluate the diagnostic performance of the International Ovarian Tumour Analysis (IOTA) logistic regression model (LR2), ultrasound Simple Rules (SR), the Risk of Malignancy Index (RMI) and subjective assessment (SA) for preoperative characterisation of adnexal masses, when ultrasonography is performed by examiners with different background training and experience.

Methods: A 2-year prospective multicentre cross-sectional study. Thirty-five level II ultrasound examiners contributed in three UK hospitals. Transvaginal ultrasonography was performed using a standardised approach. The final outcome was the surgical findings and histological diagnosis. To characterise the adnexal masses, the six-variable prediction model (LR2) with a cutoff of 0.1, the RMI with cutoff of 200, ten SR (five rules for malignancy and five rules for benignity) and SA were applied. The area under the curves (AUCs) for performance of LR2 and RMI were calculated. Diagnostic performance measures for all models assessed were sensitivity, specificity, positive and negative likelihood ratios (LR+ and LR−), and the diagnostic odds ratio (DOR).

Results: Nine-hundred and sixty-two women with adnexal masses underwent transvaginal ultrasonography, whereas 255 had surgery. Prevalence of malignancy was 29% (49 primary invasive epithelial ovarian cancers, 18 borderline ovarian tumours, and 7 metastatic tumours). The AUCs for LR2 and RMI for all masses were 0.94 (95% confidence interval (CI): 0.89–0.97) and 0.90 (95% CI: 0.83–0.94), respectively. In premenopausal women, LR2 − RMI difference was 0.09 (95% CI: 0.03–0.15) compared with −0.02 (95% CI: −0.08 to 0.04) in postmenopausal women. For all masses, the DORs for LR2, RMI, SR + SA (using SA when SR inapplicable), SR + MA (assuming malignancy when SR inapplicable), and SA were 62 (95% CI: 27–142), 43 (95% CI: 19–97), 109 (95% CI: 44–274), 66 (95% CI: 27–158), and 70 (95% CI: 30–163), respectively.

Conclusion: Overall, the test performance of IOTA prediction models and rules as well as the RMI was maintained in examiners with varying levels of training and experience.

*Correspondence: Dr A Sayasneh; E-mail: a.sayasneh@imperial.ac.uk

Received 28 January 2013; revised 12 April 2013; accepted 17 April 2013; published online 14 May 2013

Although ovarian tumours are common, most are not malignant (Menon *et al*, 2009). Correctly characterising ovarian tumours is critical, as this ensures appropriate referral of patients with cancer to specialised surgeons, which is crucial to optimise patient care and survival (Vergote *et al*, 2001; Earle *et al*, 2006; Engelen *et al*, 2006; Paulsen *et al*, 2006). By correctly recognising benign ovarian masses, conservative management may be adopted, leading to reduced morbidity while facilitating fertility preservation (Carley *et al*, 2002; Tinelli *et al*, 2006).

The most accurate way to characterise adnexal pathology is subjective assessment of ultrasound findings by experienced examiners (Timmerman *et al*, 1999; Timmerman, 2004; Valentin *et al*, 2009). However, training and experience of performing transvaginal ultrasonography varies. To mirror the test performance of experienced examiners, several ultrasound-based prediction models have been developed to help operators accurately discriminate between benign and malignant adnexal masses (Jacobs *et al*, 1990; Timmerman *et al*, 2010a; Van Holsbeke *et al*, 2012). The Risk of Malignancy Index (RMI) includes serum CA125 levels, menopausal status and ultrasound findings (Jacobs *et al*, 1990). The International Ovarian Tumour Analysis (IOTA) group developed and validated a logistic regression model (LR2) with five ultrasound parameters, which has shown excellent discrimination between benign and malignant masses (Timmerman *et al*, 2005; Timmerman *et al*, 2010b; Van Holsbeke *et al*, 2012). Furthermore, the IOTA group has described simple rules based on five ultrasound features indicating malignancy (M-features) and five features suggesting a benign lesion (B-features) (Timmerman *et al*, 2008). These rules have shown good performance on temporal and external validation (Timmerman *et al*, 2010a). A criticism of these prediction models is that they were developed and validated by experts in characterising adnexal pathology (Timmerman, 2004; Timmerman *et al*, 2005; Timmerman *et al*, 2008; Timmerman *et al*, 2010a; Timmerman *et al*, 2010b; Van Holsbeke *et al*, 2012). Accordingly, we do not know if these models maintain performance in the hands of operators with different training backgrounds and experience levels.

The primary aim of this study was to examine the performance of the IOTA LR2 model, ultrasound-based Simple Rules (SR), RMI and subjective assessment (SA) by the examiner for the preoperative characterisation of ovarian masses, when ultrasonography is performed by examiners with a range of training backgrounds and experience. We aimed to validate the performance of these approaches to the diagnosis of adnexal pathology in everyday 'real world' clinical practice.

MATERIALS AND METHODS

Study design and setting. This was a prospective multicentre cross-sectional cohort study (IOTA Phase 4B). The patients were recruited from three hospitals: two tertiary referral centres for gynaecological oncology (Queen Charlotte's and Chelsea Hospital, London (QCCH); Princess Anne Hospital, Southampton (PAH)) and one urban acute hospital partnered to Imperial College (West Middlesex University Hospital, London (WMUH)). The study was approved as an assessment of 'service improvement' by the local Joint Research Office at Imperial College Academic Health Science Center and the Research and Development Department at Southampton University Hospitals. Accordingly, no formal ethical approval was required. The guidelines of the STARD (Standards for Reporting of Diagnostic Accuracy) initiative were used (Bossuyt *et al*, 2003).

Patients were recruited consecutively from September 2010 to September 2012 at QCCH, February 2012 to September 2012 at WMUH, and May 2012 to September 2012 at PAH. All ultrasound examiners attended a half-day theoretical induction session where the ultrasound features of the rules and models used in the study were illustrated. None of the examiners were considered specialist 'experts' (level III) in performing ultrasound examinations of the ovary (EFSUMB, 2006; RCR, 2012).

Patient population and data collection. The inclusion criteria were patients presenting with at least one adnexal mass that underwent transvaginal ultrasonography at one of the participating centres. In the event of bilateral adnexal masses, the mass with the most complex ultrasound morphology was included (Timmerman *et al*, 2000, 2010b). If both masses had similar ultrasound morphology, the largest mass, or the one most easily accessible by ultrasonography was included (Timmerman *et al*, 2010b).

The exclusion criteria were (i) pregnancy, (ii) patients examined by a consultant with a special interest in gynaecological ultrasound, (iii) refusal of transvaginal ultrasonography, (iv) cytology rather than histology as an outcome, and (v) failure to undergo surgery within 120 days of the ultrasound examination.

At QCCH, a secure electronic data-collection system was developed for the study (Astraia Software, Munich, Germany). A unique identifier was generated automatically for each patient's record. Dedicated data collection forms were used for WMUH and PAH. Data security was ensured following the NHS Caldecott report guidelines (The Caldicotte Committee, 1997). Recorded clinical variables included age, current pregnancy (yes, no), and menopausal status. Women $\geqslant 50$ years who had undergone hysterectomy were defined as postmenopausal.

Transvaginal ultrasonography was performed in the standardised manner previously published by the IOTA collaboration (Timmerman *et al*, 2000; Timmerman *et al*, 2010b). Transabdominal ultrasonography was performed if a large mass could not be fully assessed transvaginally (Timmerman *et al*, 2010b). Subjective assessment of the ultrasound findings was used to classify the masses as malignant or benign. Borderline tumours were considered malignant. RMI, LR2 and SR were applied centrally and checked by statisticians at the end of the study.

Operator experience was quantified by four variables using the operator's first patient recruitment date as a reference point for time: number of years of gynaecological scanning, number of gynecology scans performed, number of ovarian masses examined, and background training (sonographer or medical doctor (MD)).

Prediction models. The logistic regression model LR2 uses six variables: (1) patient age (years); (2) presence of ascites (yes = 1, no = 0); (3) presence of blood flow within a papillary projection (yes = 1, no = 0); (4) maximal diameter of the solid component (expressed in mm and truncated at 50 mm); (5) irregular internal cyst walls (yes = 1, no = 0); and (6) presence of acoustic shadows (yes = 1, no = 0). The logistic regression model LR2 estimates the probability of malignancy for an adnexal tumour as $1/(1+\exp(-z))$, where $z = -5.3718 + 0.0354(1) + 1.6159(2) + 1.1768(3) + 0.0697(4) + 0.9586(5) - 2.9486(6)$. A probability cutoff of 0.1 (10%) was used to classify patients as benign or malignant based on LR2 (Timmerman *et al*, 2005; Timmerman *et al*, 2010b; Van Holsbeke *et al*, 2012).

The SR are based on five ultrasound features of malignancy (M-features) and five ultrasound features suggestive of a benign lesion (B-features) (Timmerman *et al*, 2008; Timmerman *et al*, 2010a). An ovarian mass is classified as malignant if at least one M-feature and no B-features are present and vice versa (Timmerman *et al*, 2008; Timmerman *et al*, 2010a). When no B- or M-features are present or if both B- and M-features are present, then SR are considered inconclusive (uncertain) and a different diagnostic method should be used (Timmerman *et al*, 2008; Timmerman *et al*, 2010a). For SR, two approaches were used: one where all inconclusive cases were classified as malignant to limit the number of missed cancers (SR + MA), and another where

inconclusive cases were classified as benign or malignant using SA by the examiner (SR + SA).

Measurements of serum CA125 were carried out according to each centre's normal practice, using Abbott Architect CA125 II (Abbott Park, IL, USA) immunoassay kit at QCCH, Advia Centaur XP Immunoassay System (Centaur) (Siemens Healthcare Diagnostics Inc., Deerfield, IL, USA) at WMUH and UniCel DxI Immunoassay System (Beckman Coulter Inc., Brea, CA, USA) Assay at PAH.

For the RMI, five features were incorporated into the ultrasound score (U): multilocularity, solid areas, bilateral masses, ascites and evidence of metastases. U was assigned a value of 0 when none of these features was present, 1 if one feature was present and 3 if two or more features were present. A score (M) of 1 was assigned to premenopausal and 3 to postmenopausal women. Risk of Malignancy Index was defined as $U \times M \times$ (serum CA125 ($U\,ml^{-1}$)). An RMI score of $\geqslant 200$ was used as the cutoff value to indicate cancer (Jacobs *et al*, 1990).

Reference standard. The final outcome was the surgical findings and histological diagnosis of removed tissues, and the classification of these as benign or malignant. Borderline tumours were classified as malignant tumours. Surgery was performed by laparoscopy or laparotomy, according to the surgeon's judgment. Excised tissues underwent histological examination at the local Department of Pathology. Tumours were classified using the criteria recommended by the International Federation of Gynecology and Obstetrics (Heintz *et al*, 2006).

Statistical analysis. For LR2 and RMI, receiver-operating characteristic curves were derived and summarised using the area under the curve (AUC) with 95% confidence interval (CI) using the logit transform method (Pepe, 2003). Since 70% of the patients were collected at one hospital, we report AUCs computed on the whole sample instead of performing a random effects meta-analysis of hospital-specific AUCs, the results of which were nearly identical. For SR, the classification has essentially three ordinal levels: benign, inconclusive (uncertain), or malignant. A receiver-operating characteristic curve for this classification was computed, which has two points (one for benign *vs* inconclusive/malignant, and one for benign/inconclusive *vs* malignant). This was done to allow a visual comparison of the performance of SR with RMI and LR2. However, an AUC was not derived because this would not be comparable to AUCs of models that give continuous results. In addition, we explored whether the performance of LR2 and SR differed from the performance of RMI. We examined the performance in pre- and postmenopausal women separately. For differences in AUC, the method of DeLong *et al* (1988) was used to generate the 95% CI.

Diagnostic performance measures were computed for the classification as benign or malignant based on RMI, LR2, SR and SA. Reported diagnostic performance measures were sensitivity, specificity, positive and negative likelihood ratios (LR+ and LR−), and the diagnostic odds ratio.

Missing CA125 levels ($n = 19$) were statistically imputed using predictive mean matching regression. Owing to heavy skewness,

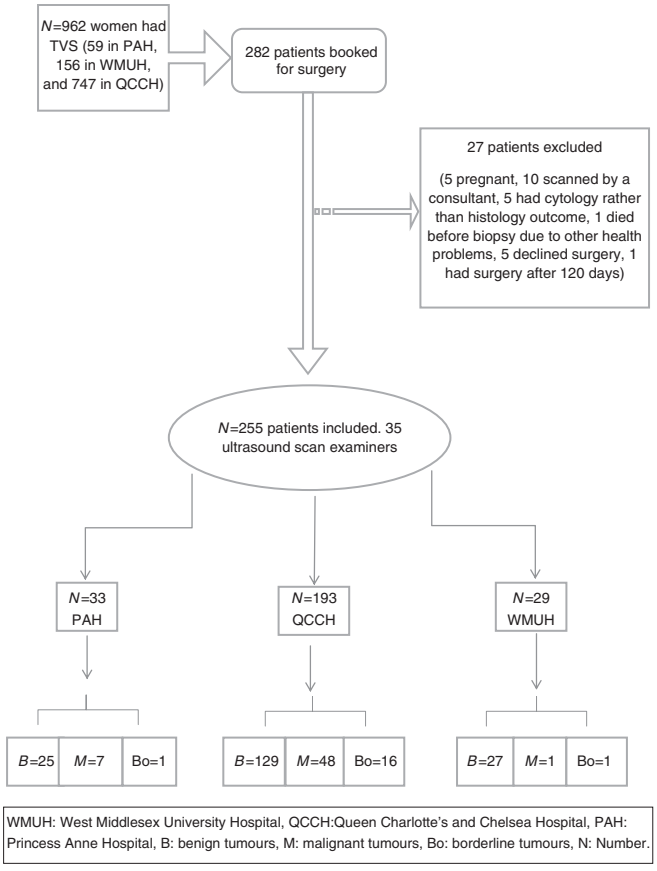

Figure 1. A flow chart illustrating the final sample size and the numbers of excluded cases.

Table 1. Different histological outcomes of ovarian lesions in the study

Type of ovarian mass on histology	N	%
Simple cyst	6	2.4
Endometrioma	39	15.3
Mature teratoma	30	11.8
Hydrosalpinx	3	1.2
Tubo-ovarian abscess/infection	5	2
Hemorrhagic cyst	8	3.1
Ovarian torsion with no histological diagnosis[a]	2	0.8
Functional cyst	12	4.7
Simple para-ovarian cyst	5	2
Ovarian fibroma	7	2.7
Serous cystadenoma	38	14.9
Mucinous cystadenoma	16	6.3
Other benign tumours[b]	10	3.9
Serous borderline tumours	8	3.1
Mucinous borderline tumours	8	3.1
Other borderline tumours[c]	2	0.8
Serous cyst/adenocarcinoma	26	10.2
Mucinous cyst/adenocarcinoma	7	2.7
Endometrioid carcinoma	6	2.4
Clear cell carcinoma	5	2
Other malignant tumours[d]	12	4.7
Total	255	100.0

[a]These two cases were confirmed at laparoscopy and two follow-up visits with transvaginal ultrasound scans over 6 months showing normal size and morphology for the ovaries after de-torsion.
[b]One Brenner tumour, four cases of Struma Ovarii, one chronic tubal pregnancy (with a negative pregnancy test), one mesenteric cyst, one fibrothecoma, one serous cystadenofibroma and one mixed mucinous cystadenoma and Brenner tumour.
[c]One borderline endometrioid tumour and one borderline mixed serous endometrioid tumour.
[d]One granulosa cell tumour, one transitional cell tumour, one signet ring cell adenocarcinoma, one peritoneal serous adenocarcinoma, five gastrointestinal adenocarcinomas, one malignant mixed Mullerian tumour (MMMT), one large cell neuroendocrine carcinoma, and one endocrine tumour.

the double log of CA125 is predicted using variables used in the prediction models, tumour pathology groups (Van Calster *et al*, 2011) and hospital (QCCH, WMUH, PAH). In all ($n = 19$) these cases, the RMI value was zero irrespective of the CA125 level, as the ultrasound score was zero.

We conducted an exploratory analysis of the influence of experience on the performance of subjective impression and the prediction models. A regression model for accuracy of subjective impression or a model was fitted using the number of ovarian mass scans (7 ordinal categories; <100, 100–200, 200–500, 500–1000, 1000–2000, 2000–5000 and 5000–10 000), background training (sonographer or MD), and tumour outcome (benign or malignant) as predictors. Outcome was added to adjust the effects of the predictors. A mixed effects model was used to account for the clustering of patients within operators. All analyses were performed using SAS 9.3 (SAS Institute, Cary, NC, USA).

RESULTS

During the study period, 962 women with an adnexal mass underwent ultrasonography and 282 of these patients were managed surgically. Twenty-seven cases were excluded: five because of pregnancy, ten were examined by a senior consultant (level III scan), five had cytology rather than histology as a final outcome, one died before surgery, five patients declined surgery and one patient had surgery >120 days of the index ultrasound scan (Figure 1). Five cases were included where histology was not available. Two cases of ovarian torsion were confirmed at laparoscopy and de-torted. The ovaries were normal in size and morphology on two follow-up ultrasound scans 3 and 6 months after the procedure. A further three cases were included where an abscess was diagnosed surgically and confirmed by microscopy and culture. The mean age of the patients was 46 years (95% CI: 34–57). One-hundred and sixty-five patients (65%) were premenopausal. The prevalence of malignancy was 29% (74 malignancies *vs* 181 benign ovarian tumours). The 74 malignancies included: 49 primary invasive epithelial ovarian cancers, 18 borderline ovarian tumours, and 7 metastatic tumours (Table 1).

For the whole study population, the diagnostic odds ratio for LR2, RMI, SR + SA, SR + MA and SA were 62 (95% CI: 27–142), 43 (95% CI: 19–97), 109 (95% CI: 44–274), 66 (95% CI: 27–158) and 70 (95% CI: 30–163), respectively (Table 2). Overall, our data suggested a significantly higher AUC for LR2 compared with RMI: 0.94 and 0.90, respectively, with an LR2 − RMI difference of 0.04 (95% CI: 0.01–0.07) (Table 2 and Figure 2). The difference in AUC between LR2 and RMI was greatest in premenopausal women (AUCs of 0.92 and 0.83 for LR2 and RMI, respectively, with a difference of 0.09, 95% CI: 0.03–0.15) but little difference was observed in postmenopausal patients (0.90 and 0.92, respectively, difference −0.02, 95% CI −0.08 to 0.04; Table 2, Supplementary Figures S1 and S2). The AUCs for discrimination between benign and borderline tumours were 0.86 (95% CI: 0.75–0.97) for LR2 and 0.77 (95% CI: 0.64–0.89) for RMI. The AUCs for discrimination between benign tumours and stage I invasive cancers were 0.94 (95% CI: 0.88–1.00) for LR2 and 0.91 (95% CI: 0.84–0.99) for RMI (Supplementary Table A).

The SR were able to classify 83.9% ($n = 214$) of the masses as benign or malignant. Of the 41 tumours where the IOTA SR were uncertain, 20 were benign and 21 malignant. When SR were able to

Table 2. Sensitivity, specificity, LR+, LR−, DOR and AUC for diagnostic models in the whole sample, premenopausal group, and postmenopausal group

	Sensitivity	Specificity	LR+	LR−	DOR	AUC
LR2						
Total sample	0.88 (0.78, 0.93)	0.90 (0.84, 0.93)	8.37 (5.49, 12.98)	0.14 (0.07, 0.24)	61.58 (26.67, 141.83)	0.94 (0.89, 0.97)
Premenopausal	0.82 (0.64, 0.92)	0.96 (0.92, 0.98)	22.51 (9.70, 53.23)	0.19 (0.08, 0.37)	121.44 (33.23, 444.27)	0.92 (0.79, 0.97)
Postmenopausal	0.91 (0.80, 0.97)	0.68 (0.53, 0.80)	2.87 (1.93, 4.61)	0.13 (0.05, 0.31)	22.50 (6.90, 72.32)	0.90 (0.82, 0.95)
RMI						
Total sample	0.72 (0.60, 0.81)	0.94 (0.90, 0.97)	12.96 (7.11, 24.03)	0.30 (0.20, 0.42)	43.16 (19.25, 96.58)	0.90 (0.83, 0.94)
Premenopausal	0.54 (0.36, 0.70)	0.96 (0.92, 0.98)	14.68 (5.97, 36.16)	0.48 (0.31, 0.67)	30.46 (9.76, 94.44)	0.83 (0.67, 0.92)
Postmenopausal	0.83 (0.69, 0.91)	0.89 (0.76, 0.95)	7.27 (3.38, 16.90)	0.20 (0.10, 0.35)	37.05 (11.27, 121.23)	0.92 (0.83, 0.96)
SR + MA[a]						
Total sample	0.91 (0.82, 0.95)	0.87 (0.82, 0.91)	7.13 (4.89, 10.58)	0.11 (0.05, 0.21)	65.75 (27.24, 157.95)	NA
Premenopausal	0.86 (0.69, 0.94)	0.88 (0.83, 0.93)	7.34 (4.54, 11.95)	0.16 (0.06, 0.36)	45.38 (14.36, 141.60)	NA
Postmenopausal	0.93 (0.82, 0.98)	0.84 (0.71, 0.92)	5.88 (3.15, 11.88)	0.08 (0.03, 0.21)	75.76 (18.87, 298.02)	NA
SR + SA[b]						
Total sample	0.86 (0.77, 0.92)	0.94 (0.90, 0.97)	15.65 (8.69, 28.76)	0.14 (0.08, 0.25)	109.44 (43.79, 273.55)	NA
Premenopausal	0.82 (0.64, 0.92)	0.96 (0.91, 0.98)	18.76 (8.66, 41.34)	0.19 (0.08, 0.37)	100.43 (28.83, 349.96)	NA
Postmenopausal	0.89 (0.77, 0.95)	0.91 (0.79, 0.96)	9.80 (4.16, 25.05)	0.12 (0.05, 0.26)	82.00 (20.93, 320.46)	NA
SA						
Total sample	0.88 (0.78, 0.93)	0.91 (0.85, 0.94)	9.35 (5.98, 14.89)	0.13 (0.07, 0.24)	69.67 (29.76, 162.80)	NA
Premenopausal	0.86 (0.69, 0.94)	0.94 (0.89, 0.97)	14.68 (7.53, 29.14)	0.15 (0.06, 0.34)	96.75 (27.65, 335.52)	NA
Postmenopausal	0.89 (0.77, 0.95)	0.80 (0.66, 0.89)	4.36 (2.54, 8.08)	0.14 (0.06, 0.30)	31.89 (9.91, 102.03)	NA

Abbreviations: AUC = area under the curve; DOR = diagnostic odds ratio; LR+ and LR− = positive and negative likelihood ratios; LR2 = logistic regression model 2 (cutoff = 0.1); MA = malignancy assumption; NA = not applicable as not a continuous numeral variable; RMI = Risk of Malignancy Index (cutoff = 200); SA = Subjective Assessment; SR = Simple Rules. Whole sample ($n = 255$, 74 malignant, 181 benign), premenopausal ($n = 165$, 28 malignant, 137 benign), and postmenopausal ($n = 90$, 46 malignant, 44 benign). Results are in value (95% CI).
[a]SR and malignancy assumption when SR are not applicable.
[b]SR and using the subjective impression when SR are not applicable.

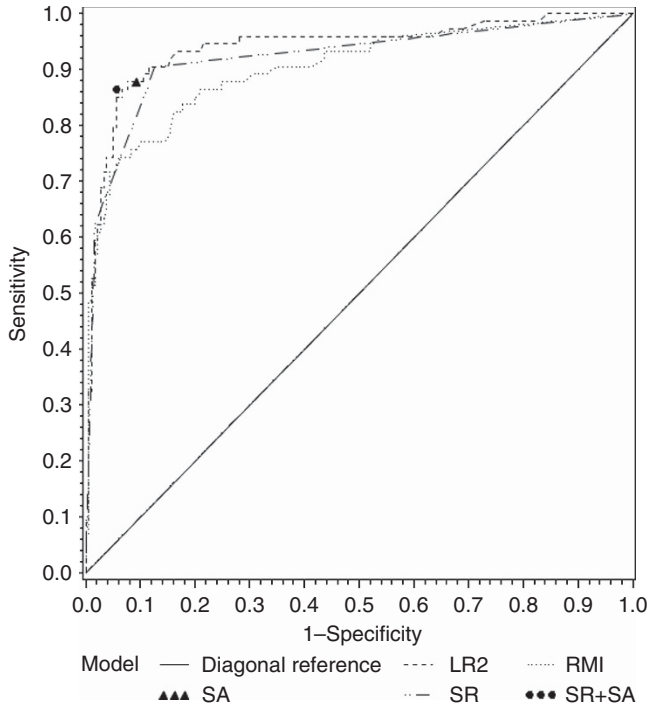

Figure 2. Receiver-operating characteristic (ROC) plot for all masses. Abbreviations: LR2 = Logistic Regression model 2; RMI = Risk of Malignancy Index; SR = Simple Rules have three levels (benign, inconclusive, and malignant) and is represented by a ROC curve with two points. SA = subjective assessment; SR + SA = SR and using SA by examiner when SR were inconclusive.

Table 3. The number of ovarian mass scans performed by operators

Number of ovarian mass scans	Number of operators	Percent
<100	5	14
100–200	7	20
200–500	7	20
500–1000	3	9
1000–2000	7	20
2000–5000	4	11
5000–10 000	2	6

characterise the ovarian mass, sensitivity was 87% (95% CI: 75–93%) and specificity was 98% (95% CI: 95–99%).

A strategy classifying all SR inconclusive tumours as malignant (SR + MA) yielded a significantly higher sensitivity (91%) than using the RMI (72%) (difference in sensitivity 0.19, 95% CI: 0.07–0.31). However, the specificity of this strategy was lower (87% *vs* 94% for SR + MA and RMI, respectively) (difference −0.07, 95% CI: −0.13 to 0.01). When examiners used their own SA as a second-stage test when SR were inconclusive, sensitivity was significantly higher than for RMI: 86% and 72%, respectively (difference 0.15, 95% CI: 0.02–0.27) with no difference in specificity.

In all, 62.9% of the operators have performed <1000 ultrasound scans (Table 3); 24% of the operators were MDs, whereas 76% were sonographers. The exploratory analysis of the influence of operator experience and training on diagnostic performance suggested that MDs were more able to subjectively assess the correct diagnosis than sonographers (odds ratio 2.59, 95% CI: 0.77–8.74) (Figure 3).

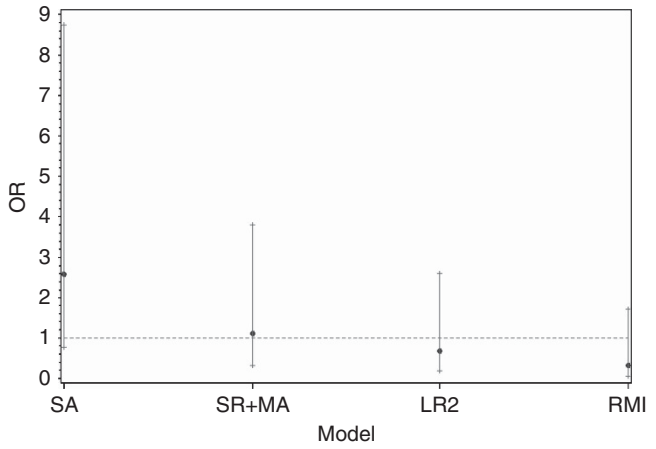

Figure 3. Plot of odds ratios OR (95% CI) of MD *vs* sonographer for each of the models. Dot: OR, line segment: 95% CI, dashed line: OR of 1 (no accuracy difference between sonographer and MD). Abbreviations: OR = odds ratio; SA = subjective assessment; SR + MA = Simple Rules and malignancy assumption when simple rules are not applicable; LR2 = Logistic Regression model 2; RMI = Risk of Malignancy Index.

When using SR + MA and LR2 to classify masses as benign or malignant, the odds ratios were 1.10 (95% CI: 0.32–3.81) and 0.68 (95% CI: 0.18–2.61, respectively, suggesting similar performance of these models in the hands of MDs and sonographers. When using the RMI, the odds ratio was 0.32 (95% CI: 0.06–1.70), suggesting slightly better performance for sonographers. The number of previous ovarian mass scans had little effect, with odds ratios between 0.85 and 1.01 for each category increase on the ordinal measurement scale. Adding hospital as a fixed effect in the mixed effects model had no influence on the final results (Supplementary Table B).

DISCUSSION

We have shown that the IOTA LR2 model and SR perform well in the hands of examiners with different background training or relatively little experience using ultrasonography. Criticism of papers describing the external validation of IOTA and other models has focused on the fact that they were developed and tested by examiners with a specific expertise in imaging of adnexal pathology (Timmerman *et al*, 2005; Timmerman *et al*, 2008; Timmerman *et al*, 2010a; Timmerman *et al*, 2010b; Van Holsbeke *et al*, 2012; Kaijser *et al*, 2013). In contrast, in the current study the ultrasound scans were performed by examiners with different training (sonographers and doctors) and level II experience. Our findings agree with the IOTA group external validation for LR2, where the AUC for LR2 was 0.94 compared with 0.90 for RMI for the whole study population (Van Holsbeke *et al*, 2012). Despite sample size limitations when stratifying for menopausal status, our results were similar to the IOTA external validation study, with LR2 offering a clear diagnostic advantage over RMI for premenopausal patients, although not in the postmenopausal group.

To our knowledge, this study represents the first external validation of the IOTA LR2 and SR by examiners with a range of experience and training; furthermore, the patients were seen in different centres. As most ovarian pathology is probably examined by sonographers or doctors who do not have a special interest in gynaecologic ultrasonography (level II), it seems reasonable to suggest that our findings offer clinicians a clearer idea on the

performance of the different adnexal mass risk models in daily practice. In 2012, Nunes *et al* (2012) externally validated the IOTA LR2 model on 124 women by a single relatively inexperienced gynaecologist (level II). They reported an AUC of 0.93 for LR2 but did not compare RMI, LR2 and SR nor stratify the AUCs according to menopausal status.

A strength of our study is that it adhered to a strict prospective protocol, took place in three units and drew on a relatively large number of examiners. A weakness common to other studies is the difficulty encountered in classifying operator experience. Similarly, when the Royal College of Radiologists in the United Kingdom published recommendations for ultrasound training for medical and surgical specialties, it found it difficult to define boundaries between the three levels of ultrasound scanning experience proposed (RCR, 2012). In our study, 67 (25.97%) patients were examined by sonographers, who are not considered in the Royal College of Radiologists recommendations (RCR, 2012). Interestingly, we found that subjective impression of the nature of an adnexal mass tended to be better by medically trained examiners than sonographers. However, this difference was not seen when doctors and sonographers were asked to enter ultrasound findings into the prediction models LR2 and RMI or when they used SR (Figure 2). This is likely to reflect variations in training, as sonographers are in general taught to identify and report the structures they see. Hence, they are skilled at accurately entering the presence or absence of the structures required for use in prediction models but are less likely to offer an opinion on the final diagnosis. This is an important observation, as the original aim of the IOTA study was to develop tools that could be used by all examiners to enhance their diagnostic performance.

In our study, there was a variation in the CA125 kits used in each centre. This slight variation was previously assessed and found to have very limited impact on the variation in diagnostic accuracy of these kits (Davelaar *et al,* 1998). Moreover, it has been suggested that the use of different CA125 assay kits reflects 'real world' clinical practice and will produce more generally applicable results (Van Calster *et al,* 2011).

An advantage to using LR2 is it provides clinicians with absolute risks of a patient having ovarian cancer, which may contribute to patient counselling and shared decision making. In clinical practice, calculating LR2 may sound more difficult to use than SR. To facilitate its use, the LR2 formula can easily be made available online, and incorporated into mobile applications or computer software (Van Belle *et al*, 2012). Our data show that overall diagnostic performance is better with LR2 compared with RMI, but also suggest that LR2 misses fewer borderline (AUC of 0.86 for LR2 *vs* 0.77 for RMI) and stage 1 invasive ovarian cancers (AUC of 0.94 for LR2 *vs* 0.91 for RMI).

In our study, SR could be applied to 83.9% of the study population compared with 77% in the original IOTA external validation (Timmerman *et al*, 2010a). The sensitivity and specificity for SR in the hands of the examiners in our study was 87% and 98%, compared with 92% and 96%, respectively, in the original IOTA study (Timmerman *et al*, 2010a). The utility of SR is supported by Fathallah *et al* (2011), who conducted a single-centre external validation study on 122 ovarian tumours over 4 years. They found SR were applicable in 89.3% of the study population, with a sensitivity of 73% and specificity of 97%. However, they did not evaluate different strategies for second-stage tests in the event of SR being inconclusive (Fathallah *et al*, 2011). Ideally, when the SR are inconclusive, the patient should be referred to an expert in gynaecological scanning for further assessment (level III) as an optimal second-stage test. In the absence of level III ultrasonography, our data suggest that if SR are inconclusive, an acceptable second-stage test for level II doctor ultrasound examiners is the subjective impression of the scan findings. For sonographers, however, a reasonable strategy would be to classify all such lesions as malignant. When SR are inconclusive, another alternative to be considered, especially when an experienced level III ultrasound examiner is not available, is to refer the patient for an MRI for these more difficult masses (Bernardin *et al*, 2012). However, further studies are needed before adopting this as a protocol.

Correctly classifying the nature of ovarian pathology is a common diagnostic problem in gynecology, and correctly identifying the presence of cancer in these cases is the key to ensure patients access appropriate treatment. This study shows that the IOTA LR2 model and SR perform well in the hands of both relatively inexperienced doctors and when used by sonographers. Furthermore, although not the primary aim of this study, our data suggest the performance of the both LR2 and SR may be better than the RMI. These findings suggest that LR2 or SR may replace the RMI in protocols designed to evaluate suspected adnexal pathology, particularly when dealing with premenopausal women.

ACKNOWLEDGEMENTS

AS is a clinical research fellow at the Imperial College London. This study was funded by the Imperial College London. The funding source had no role in the design, data collection, data analysis, or interpretation of the findings. AS and TB had full access to all the data in the study and had the final responsibility to submit for publication. LW is supported by a PhD grant of the Flanders' Agency for Innovation by Science and Technology (IWT Vlaanderen). BVC is a postdoctoral fellow of the Research Foundation—Flanders (FWO). Tom Bourne and Sadaf Ghaem-Maghami are supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at the Imperial College Healthcare NHS Trust and Imperial College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. We thank all ultrasound examiners who contributed in this study.

REFERENCES

Bernardin L, Dilks P, Liyanage S, Miquel ME, Sahdev A, Rockall A (2012) Effectiveness of semi-quantitative multiphase dynamic contrast-enhanced MRI as a predictor of malignancy in complex adnexal masses: radiological and pathological correlation. *Eur Radiol* **22**(4): 880–890.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, 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.

Carley ME, Klingele CJ, Gebhart JB, Webb MJ, Wilson TO (2002) Laparoscopy versus laparotomy in the management of benign unilateral adnexal masses. *J Am Assoc Gynecol Laparosc* **9**(3): 321–326.

DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**(3): 837–845.

Davelaar EM, van Kamp GJ, Verstraeten RA, Kenemans P (1998) Comparison of seven immunoassays for the quantification of CA 125 antigen in serum. *Clin Chem* **44**: 1417–1422.

Earle CC, Schrag D, Neville BA, Yabroff KR, Topor M, Fahey A, Trimble EL, Bodurka DC, Bristow RE, Carney M, Warren JL (2006) Effect of surgeon specialty on process of care and outcomes for ovarian cancer patients. *J Natl Cancer Inst* **98**: 172–180.

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

Engelen MJ, van der Zee AG, de Vries EG, Willemse PH (2006) Debulking surgery for ovarian epithelial cancer performed by a gynaecological

oncologist improved survival compared with less specialised surgeons. *Cancer Treat Rev* **32**(4): 320–323.

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]. *Gynecol Obstet Fertil* **39**(9): 477–481.

Heintz AP, Odicino F, Maisonneuve P, Quinn MA, Benedet JL, Creasman WT, Ngan HY, Pecorelli S, Beller U (2006) Carcinoma of the ovary. FIGO 6th Annual Report on the Results of Treatment in Gynecological Cancer. *Int J Gynaecol Obstet* **95**(Suppl 1): S161–S192.

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, Bourne T, Valentin L, Sayasneh A, Van Holsbeke C, Vergote I, Testa AC, Franchi D, Van Calster B, Timmerman D (2013) Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumour Analysis (IOTA) studies. *Ultrasound Obstet Gynecol* **41**(1): 9–20.

Menon U, Gentry-Maharaj A, Hallett R, Ryan A, Burnell M, Sharma A, Lewis S, Davies S, Philpott S, Lopes A, Godfrey K, Oram D, Herod J, Williamson K, Seif MW, Scott I, Mould T, Woolas R, Murdoch J, Dobbs S, Amso NN, Leeson S, Cruickshank D, McGuire A, Campbell S, Fallowfield L, Singh N, Dawnay A, Skates SJ, Parmar M, Jacobs I (2009) Sensitivity and specificity of multimodal and ultrasound screening for ovarian cancer, and stage distribution of detected cancers: results of the prevalence screen of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). *Lancet Oncol* **10**(4): 327–340.

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

Paulsen T, Kjaerheim K, Kaern J, Tretli S, Tropé C (2006) Improved short-term survival for advanced ovarian, tubal, and peritoneal cancer patients operated at teaching hospitals. *Int J Gynecol Cancer* **16**(Suppl 1): 11–17.

Pepe MS (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, UK.

The Royal College of Radiologists (RCR), Board of the Faculty of Clinical Radiology (2012) *Ultrasound Training Recommendations for Medical and Surgical Specialties. Royal College Radiologists*. 2nd edn (The Royal College of Radiologists: Londonhttp://www.rcr.ac.uk/docs/radiology/pdf/BFCR(12)17_ultrasound_training.pdfAccessed 25 January 2013.

The Caldicotte Committee. Chaired by Caldicott DF (1997) *Report on the review of patient-identifiable information*. (ed) Department of Health: London. http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4068404.pdf. Accessed 11 November 2012.

Timmerman D (2004) The use of mathematical models to evaluate pelvic masses; can they beat an expert operator? *Best Pract Res Clin Obstet Gynaecol* **18**(1): 91–104.

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**(dec14 1): c6839–c6839.

Timmerman D, Schwärzler P, Collins WP, Claerhout F, Coenen M, Amant F, Vergote I, Bourne TH (1999) Subjective assessment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience. *Ultrasound Obstet Gynecol* **13**(1): 11–16.

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 (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 T, Collins WP, Verrelst H, Vergote I (2000) Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumour analysis (IOTA) group. *Ultrasound Obstet Gynecol* **16**: 500–505.

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**: 226–234.

Tinelli R, Tinelli A, Tinelli FG, Cicinelli E, Malvasi A (2006) Conservative surgery for borderline ovarian tumors: a review. *Gynecol Oncol* **100**(1): 185–191.

Valentin L, Jurkovic D, Van Calster B, Testa A, Van Holsbeke C, Bourne T, Vergote I, Van Huffel S, Timmerman D (2009) Adding a single CA 125 measurement to ultrasound imaging performed by an experienced examiner does not improve preoperative discrimination between benign and malignant adnexal masses. *Ultrasound Obstet Gynecol* **34**: 345–354.

Van Belle VM, Van Calster B, Timmerman D, Bourne T, Bottomley C, Valentin L, Neven P, Van Huffel S, Suykens JA, Boyd S (2012) A mathematical model for interpretable clinical decision support with applications in gynecology. *PloS One* **7**(3): e34312.

Van Calster B, Valentin L, Van Holsbeke C, Zhang J, Jurkovic D, Lissoni AA, Testa AC, Czekierdowski A, Fischerova 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**: 815–825.

Vergote I, De Brabanter J, Fyles A, Bertelsen K, Einhorn N, Sevelda P, Gore ME, Kaern J, Verrelst H, Sjovall K, Timmerman D, Vandewalle J, Van Gramberen M, Trope CG (2001) Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma. *Lancet* **357**(9251): 176–182.

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