Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis

Jeroen Kaijser¹,²†, Ahmad Sayasneh³†, Kirsten Van Hoorde⁴, Sadaf Ghaem-Maghami³, Tom Bourne¹,²,³, Dirk Timmerman¹,²* and Ben Van Calster¹

¹Department of Development and Regeneration, KU Leuven, Leuven, Belgium ²Department of Obstetrics and Gynecology and Leuven Cancer Institute, University Hospital KU Leuven, Herestraat 49, Leuven 3000, Belgium ³Department of Surgery and Cancer, Queen Charlotte’s & Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, UK ⁴Department of Electrical Engineering (ESAT) and iMinds Future Health Department, KU Leuven, Leuven, Belgium

Correspondence address. Tel: +32-16-344201; Fax: +32-16-344216; E-mail: dirk.timmerman@uzleuven.be

Submitted on August 9, 2013; resubmitted on October 28, 2013; accepted on November 5, 2013

TABLE OF CONTENTS

- Introduction
- Methods
  - Protocol and registration
  - Eligibility criteria
  - Search and information services
  - Data collection process
  - Data extraction and quality assessment
  - Statistical analysis and quantitative data synthesis
- Results
  - Study selection, study characteristics and risk of bias within studies
  - Synthesis of results
- Discussion
  - Summary of evidence
  - Strengths and limitations
  - Unanswered questions and future research
- Conclusion

BACKGROUND: Characterizing ovarian pathology is fundamental to optimizing management in both pre- and post-menopausal women. Inappropriate referral to oncology services can lead to unnecessary surgery or overly radical interventions compromising fertility in young women, whilst the consequences of failing to recognize cancer significantly impact on prognosis. By reflecting on recent developments of new diagnostic tests for preoperative identification of malignant disease in women with adnexal masses, we aimed to update a previous systematic review and meta-analysis.

† The authors consider that the first two authors should be regarded as joint First Authors.

© The Author 2013. Published by Oxford University Press on behalf of the European Society of Human Reproduction and Embryology. All rights reserved.
For Permissions, please email: journals.permissions@oup.com
## Introduction

Detecting ovarian cancer in primary care is challenging because of the nonspecific nature of the symptoms. Once women have reached secondary care, characterizing ovarian pathology is fundamental to optimizing patient management (Vergote et al., 2001; Bristow and Berek, 2006; Woo et al., 2012). Ovarian tumours are a common clinical entity that affects women of all ages (Alcazar et al., 2005). It has been estimated that in the USA a woman has a 5–10% lifetime risk of undergoing surgery for this condition (Curtin, 1994). Although, a priori, most tumours are benign, it is of paramount importance to correctly characterize whether ovarian masses constitute benign or malignant disease. This is particularly the case in premenopausal women where preservation of fertility is an important issue. Although these women are rarely thought to have cancer, they in fact account for up to 20% of all ovarian malignancies (Bristow et al., 2004; Siegel et al., 2011).

Patients with diagnosed ovarian cancer should be treated in specialist centres that provide the most comprehensive cancer care (Woo et al., 2012), whereas for benign ovarian masses expectant or conservative surgical management may be warranted due to reduced morbidity and the importance of fertility preservation (Carley et al., 2002; Tinelli et al., 2006). Even in selected cases of early stage invasive disease or borderline ovarian tumours (BOT), conservative treatment is a therapeutic option in young women who want to preserve their childbearing capacity (Morice et al., 2003; Tinelli et al., 2006; Darai et al., 2013). Moreover, since ‘oncogility’ was introduced in 2006 as a well-identified subspecialty that focuses on cancer survivors, ovarian cancer characterization has become more important for exploring all fertility preservation options before any cancer surgery (Woodruff, 2007).

Unnecessary surgery or overly radical surgical interventions with abrupt loss of childbearing potential in young women are all significant risks to patients with a cyst that is inappropriately characterized, whilst the consequences of failing to recognize a cyst as malignant significantly impact on prognosis (Vergote et al., 2001; Fanfani et al., 2004; Yazbek et al., 2008).

Prediction models have been developed to assist clinicians to triage patients to appropriate treatment pathways; however, none has gained universal acceptance in routine daily practice. For example in the UK, the National Institute of Clinical Excellence and Royal College of Obstetrics and Gynaecology guidelines suggest using the Risk of Malignancy Index (RMI) as a primary diagnostic tool in secondary care, whereas the American Congress of Obstetrics and Gynaecology recommends using a combination of clinical, demographic, laboratory and imaging variables (Dearking et al., 2007; RCOG guideline, green top 34, 2010; NICE clinical guidelines CG122, 2011). The current test performance of these algorithms depends heavily upon biomarkers (serum CA 125 levels), and this limits their utility, particularly in women of reproductive age (Kaijser et al., 2013b). Serum CA 125 levels are frequently normal in borderline tumours and early stage invasive ovarian cancer (Engelen et al., 2000; Gotlieb et al., 2000), and can show a false-positive increase in numerous benign tumours or conditions that irritate the pelvic peritoneum (e.g. endometriosis, fibroids, pregnancy, infection and surgery) (Sevinc et al., 2007). More recently, other biomarkers (Moore et al., 2007) have been combined into multi-marker decision algorithms such as the Risk of Ovarian Malignancy Algorithm (ROMA) and OVA-1 to triage patients in both primary and secondary care (Bast et al., 2012). Other diagnostic models have been only based on clinical information and ultrasound features. The International Ovarian Tumour Analysis (IOTA) models and rules (LR2 and Simple Rules) characterize adnexal tumours based upon the presence or absence of typical ultrasound features of malignancy (e.g. ascites, increased vascularization, solid components, tumour size, papillary projections and irregular cyst walls) (Timmerman et al., 2005, 2008).
Recent systematic reviews that evaluate the diagnostic test performance of several prediction models have only reviewed the literature up to 2009 (Geominì et al., 2009; Dodge et al., 2012). As a result, numerous validation studies assessing the ROMA and the ultrasound-based models and rules (LR2 and Simple Rules) of the largest study on the classification of ovarian pathology in the literature to date (IOTA) have not been subject to detailed review. Therefore, the conclusions of these earlier reviews have important limitations and are, by definition, outdated.

The objective of this study was to conduct an up-to-date systematic review and meta-analysis that reflects the recent development of new diagnostic tests for the preoperative identification of malignant disease in women with adnexal masses.

**Methods**

**Protocol and registration**

All methods for analysis, inclusion/exclusion criteria, data extraction and quality assessment were specified in advance. The systematic review and meta-analysis was performed in accordance with recommended methods (Irwig et al., 1994; Deeks, 2001; Moher et al., 2009; Macaskill et al., 2010; http://www.prisma-statement.org/).

**Eligibility criteria**

Eligible studies for qualitative data synthesis had to contain sufficient data to extract 2 × 2 contingency tables on the diagnostic test performance of prediction models that use at least two parameters (e.g., diagnostic imaging variables, biomarkers, patient demographics, etc.) to estimate the risk of malignancy in women presenting with an adnexal mass. Studies fulfilling the following criteria were excluded: studies limited to children, adolescents or pregnant women; studies that evaluated model performance for only very specific histological subgroups of ovarian cancer (i.e., non-epithelial ovarian cancer, BOTs); studies that evaluated technical aspects of a prediction model rather than its diagnostic accuracy; studies that are reported only in abstract form and either the full text or the raw data from the study investigators were not available; studies that did not evaluate prediction models for ovarian cancer, or evaluated only models containing one parameter; studies that used duplicate data; studies that tested models in a screening setting; studies that used healthy subjects as controls; and studies in which we were unable to extract 2 × 2 tables cross-classifying the model results and histopathology. No publication status or study design restrictions were imposed.

**Search and information sources**

To identify eligible studies, searches were run in the following electronic databases: MEDLINE (PubMed) and EMBASE (OvidSP) from March 2008 to the date of the last search (1 October 2013). The search strategy included exactly the same search terminology for both the index test and target condition as previously described (Geominì et al., 2009). The MEDLINE search was performed using the following keywords: [‘ovarian neoplasms’[MeSH] NOT ‘therapeutics’[MeSH] AND ‘model’] and [‘ovarian neoplasms’[MeSH] NOT ‘therapeutics’[MeSH] AND ‘prediction’]. The EMBASE search was performed using the following keywords: [ovary tumor AND prediction], [ovary tumor AND Mathematical model] and [ovary tumor AND statistical model]. Additional eligible articles were identified by manually searching cross references of articles retrieved from the computerized databases and relevant reviews.

**Data collection process**

Two assessors (J.K. and A.S.) independently reviewed all citations identified by the search strategy described above, first by title and abstract and if necessary by reviewing the full text of the study report, to determine eligibility. To be included in the qualitative data synthesis, the full text of each potentially relevant article was obtained and assessed to determine whether the study met the inclusion and exclusion criteria detailed above. Any disagreement was resolved by consensus. For excluded studies, reasons for exclusion were documented. The references of these excluded studies (n = 50) can be found as supporting information to this review (Supplementary data, Text S1).

**Data extraction and quality assessment**

Both reviewers (J.K. and A.S.) independently completed a predefined data extraction sheet as discussed below whilst discrepancies were resolved by consulting a third party (B.V.C., T.B. or D.T.). Information from articles in languages other than English, Dutch, French or German was only extracted when an English abstract and required data were available. The following data were retrieved:

- General information: author, journal, year of publication, cross-sectional versus case—control study design, prospective versus retrospective data collection, multicentre design and sampling method.
- Total sample size of included adnexal masses, prevalence of malignancy and allocation of BOTs to benign or malignant subgroup.
- Type of prediction model validated, its variables and cut-off used for test positivity.
- Type of reference standard used.
- Number of true positives (TPs), true negatives (TNs), false positives, false negatives. If these counts were not reported, contingency tables were reconstructed from summary estimates of sensitivity and specificity if provided. In the event that studies reported sensitivity/specificity pairs for different cut-off values, these were all registered. AUCs were also retrieved if available with their corresponding 95% confidence intervals.
- Methodological quality items and types of bias were evaluated using a modified version of the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) assessment tool. This is a validated checklist for the assessment of the methodological quality of studies included in systematic reviews of diagnostic accuracy developed by the NHS Centre for Reviews and Dissemination, University of York, UK (Whiting et al., 2003).

**Statistical analysis and quantitative data synthesis**

Meta-analysis was restricted to those prediction models that had been externally validated in at least two different studies on a minimum total sample of 1000 adnexal masses in order to retrieve more robust, generalizable results. Additionally, for eligible multicentre studies we contacted the corresponding authors to retrieve overall centre-specific results if these were not provided in the report. When we were not able to obtain centre-specific results, we continued with the reported pooled result.

All prediction models included in the meta-analysis were developed to predict malignancy in women with any adnexal tumour, except for the ROMA algorithm, which more specifically differentiates between epithelial ovarian cancer and benign disease and so previous validation studies have systematically excluded non-epithelial malignant ovarian tumours, metastatic disease, non-ovarian primary tumours or BOTs (Bandiera et al., 2011; Kim et al., 2011; Molina et al., 2011; Montagnana et al., 2011; Ruggeri et al., 2011; Kadija et al., 2012; Karlsen et al., 2011; Partheen et al., 2011; Chan et al., 2013; Sandri et al., 2013). In order to achieve a more valid comparison between all models, we decided to only use results for those validation studies of ROMA (Moore et al., 2009, 2011; Jacob et al., 2011; Lenhard et al., 2011; Kadija et al., 2012; Karlsen et al., 2011; Partheen et al., 2011; Chan et al., 2013; Sandri et al., 2013).
et al., 2011; Van Gorp et al., 2011; Anton et al., 2012; Presl et al., 2012; Pitta et al., 2013) that have reported diagnostic test performance for all types of adnexal malignancy. The same held true for combining dichotomized levels (elevated versus normal) of the individual biomarkers serum CA 125 and HE4, which was regarded as a ‘model’ according to our eligibility criteria. Although it met the predefined criteria for meta-analysis, we decided to exclude it from further analysis as only one validation study (Chang et al., 2011) reported overall test performance for this model without prior exclusion of certain histological subgroups.

In case we detected statistical inconsistencies that we were unable to clarify with the corresponding author of the original article, we excluded the study from the meta-analysis. Validation studies of prediction models that excluded BOTs or allocated BOTs to the benign cohort were not used in the quantitative data synthesis.

All individual study results were presented graphically by plotting estimates of sensitivity and specificity and their 95% confidence intervals to display the precision of the estimates and to illustrate heterogeneity between studies. For models where studies used a common cut-off point to define a positive test results, a pooled summary estimate of the expected operating point (sensitivity and specificity) and corresponding 95% confidence interval was obtained using a bivariate random effects model using the PROC GLIMMIX procedure in SAS 9.2 (SAS Institute, Cary, NC, USA) (Macaskill et al., 2010). If there was variation in the cut-off points used to define a positive test, a bivariate hierarchical summary receiver operating curve (HSROC) model was fitted using the PROC NLMIXED procedure in SAS 9.2 to estimate the summary ROC curve (Macaskill et al., 2010).

Although we intended to perform a subgroup analysis by menopausal status for all included models, this was not feasible as the majority of validation studies did not stratify their results post hoc for menopausal status. However, we had individual patient data available for pre- and post-menopausal women from two multicentre cohorts (Timmerman et al., 2010; Van Holsbeke et al., 2012; Sayasneh et al., 2013b) on which several models were evaluated, so we conducted a meta-analysis for both groups on these data to minimize the risk of introducing bias. For both cohorts, we used centre-specific data from 15 units in different clinical settings involving a total number of 1252 patients. Only units that contributed at least 3 benign and 3 malignant cases were considered in this analysis.

## Results

### Study selection, study characteristics and risk of bias within studies

Figure 1 summarizes the process of identification and selection of studies. We retrieved a total (n = 82) of 79 new research articles and three systematic reviews (Dodge et al., 2012; Li et al., 2012; Ferraro et al., 2013) in addition to the research articles contained (n = 116) in the previous systematic review (Geomini et al., 2009). These 195 validation studies have reported on 116 different prediction models for differentiating between the benign or malignant nature of an adnexal mass. Our extended search retrieved 33 new models in addition to the 83 models already derived from the previous review (Geomini et al., 2009).

Figure 2 shows a summary of the main quality characteristics of the 195 included studies according to the modified version of the QUADAS assessment tool. Overall 79% of the included studies were cross-sectional (‘cohort-type’), 59% were prospective and 55% sampled consecutive women. Most studies (82%) included patients with an adnexal mass that underwent both the index test(s) and a reference standard in order to avoid verification bias. The vast majority of studies (93%) used histology as their final outcome, which in general is considered to be an acceptable reference standard. However, 63% did not report on the time interval between the index and reference test, 51% did not report withdrawal and finally 63% did not comment on blinding of the index test results.

### Synthesis of results

In total 19 prediction models met our predefined criteria for further meta-analysis. Table 1 displays all models considered in the review with their parameters, and their total number of validation studies. Seven of them (Sassone, Lerner, Depriest, Ferrazzi, Tailor, RMI I to II) had already been extensively evaluated in the previous systematic review (Jacobs et al., 1990; Sassone et al., 1991; DePriest et al., 1993; Lerner et al., 1994; Tingulstad et al., 1996; Ferrazzi et al., 1997; Tailor et al., 1997). In contrast, we evaluated an additional 12 models (Prömpeler et al., 1997; Timmerman et al., 1999a, b, 2008; Tingulstad et al., 1999; Jukubikiené et al., 2007; Moore et al., 2009; Van Holsbeke et al., 2009; Yamamoto et al., 2009; Ueland et al., 2011). All 19 models combined had been tested on a total of 26 438 adnexal tumours including 7199 (27%) malignant and 19 239 (73%) benign tumours in 96 external validation studies (Davies et al., 1993; Schneider et al., 1993; Timor-Tritsch et al., 1993; DePriest et al., 1994; Sengoku et al., 1994; Botta and Zarcone, 1995; Hagen et al., 1995; Caruso et al., 1996; Leeners et al., 1996; Predanic et al., 1996; Tingulstad et al., 1996; Ferrazzi et al., 1997, 2005; Buckshee et al., 1998; Clayton et al., 1999; Morgante et al., 1999; Valentin, 1999; Aslam et al., 2000a, b; Valentin, 2000; Alcazar and Lopez-Garcia, 2001; Alcazar et al., 2001, 2003; 2013; Durdevic, 2001; Gramellini et al., 2001; Manjunath et al., 2001; Mol et al., 2001; Valentin et al., 2001; Wanapirak et al., 2001; Berlanda et al., 2002; Torres et al., 2002; Andersen et al., 2003; Itakura et al., 2003; Ma et al., 2003; Asif et al., 2004; Obeidat et al., 2004; Szpurek et al., 2005; Topuz et al., 2005; Romagnolo et al., 2006; Temple et al., 2006; Yazbek et al., 2006; Daponte et al., 2007; Van Holsbeke et al., 2007; Ulusoy et al., 2007; Chia et al., 2008; Kalghatgi-Kulkarni and Kushtagi, 2008; Clarke et al., 2009; Enakpene et al., 2009; Harry et al., 2009; Mansour et al., 2009; Moothiha and Yuenyao, 2009; Moore et al., 2009, 2011; Yamamoto et al., 2009; Desai et al., 2010; Kader Ali Mohan et al., 2010; Lou et al., 2010; Smoleń et al., 2010; Timmerman et al., 2010; Van den Akker, 2010; Aktürk et al., 2011; Ashrafangooei and Rezaeazadeh, 2011; Bandiera et al., 2011; Bouzari et al., 2011; Fathallah et al., 2011; Jacob et al., 2011; Kim et al., 2011; Lenhard et al., 2011; Molina et al., 2011; Montagnana et al., 2011; Partheen et al., 2011; Radosa et al., 2011; Rossi et al., 2011; Ruggeri et al., 2011; Ueland et al., 2011; Vaes et al., 2011; Van den Akker et al., 2011; Van Gorp et al., 2011; Anton et al., 2012; Håkansson et al., 2012; Hartman et al., 2012; Kadija et al., 2012; Karlsten et al., 2012; Nunes et al., 2012; Presl et al., 2012; Vaes et al., 2012; Van Gorp et al., 2012; Van Holsbeke et al., 2012; Chan et al., 2013; Ong et al., 2013; Pitta et al., 2013; Sandri et al., 2013; Sayasneh et al., 2013b; Terzic et al., 2013; Bristow et al., 2013).

A limited number of these studies simultaneously validated more than one model on the same dataset (Caruso et al., 1996; Ferrazzi et al., 1997; Morgante et al., 1999; Aslam et al., 2000b; Alcazar et al., 2001, 2003; Gramellini et al., 2001; Manjunath et al., 2001; Mol et al., 2001; Szpurek et al., 2005; Van Holsbeke et al., 2007; Kalghatgi-Kulkarni and Kushtagi, 2008; Clarke et al., 2009; Moothiha et al., 2009; Yamamoto et al., 2009; Kader Ali Mohan et al., 2010; Aktürk et al., 2011; Bouzari...
et al., 2011; Anton et al., 2012; Lou et al., 2010; Vaes et al., 2012; Van Holsbeke et al., 2012; Karlsen et al., 2012; Ong et al., 2013; Sayasneh et al., 2013b). All models were categorized based on model type [i.e. morphological scoring system, RMI and variants (multimodal scoring system), logistic regression models, artificial neural networks, ultrasound rules and ROMA (biomarker algorithm)].

Pooled summary estimates of the expected operating point (sensitivity and specificity) at their original cut-off point and corresponding 95% confidence interval for the considered models are shown in Table II. In contrast with the other models, ROMA considers different cut-off values for pre- and post-menopausal patients. These values also change according to the assay-kits being used to measure both serum HE4 and CA125. Although, the eight available validation studies of ROMA (Moore et al., 2009, 2011; Jacob et al., 2011; Lenhard et al., 2011; Van Gorp et al., 2011; Anton et al., 2012; Presl et al., 2012; Pitta et al., 2013) used different kits, there is little variation in sensitivity and specificity (Fig. 3) except for one study (Pitta et al., 2013) that showed a remarkable low overall sensitivity of 56.7%. Therefore it seemed redundant to fit an HSROC curve, and we were confronted with computational problems when doing so. The multivariate index assay OVA-1 also uses different cut-off values for pre- and post-menopausal women and has been validated by two large multicentre studies (Ueland et al., 2011; Bristow et al., 2013) in the USA (Table I). OVA-1 achieved comparable diagnostic accuracy in both studies with similar sensitivities of 92 and 93% and corresponding specificities of 54 and 43%. Forest plots of the individual study results for the other models are shown in Supplementary data, Figs S1–S5.

Figure 4 represents the overview of all summary point estimates and 95% CIs included in this meta-analysis together with a HSROC curve for RMI-1. The results indicate that the IOTA LR2 model (sensitivity 0.92: 95% CI 0.88–0.95; specificity 0.83: 95% CI 0.77–0.88) and the IOTA SR (sensitivity 0.93: 95% CI 0.89–0.95; specificity
0.81: 95% CI 0.76–0.85) achieve the highest diagnostic accuracy to differentiate between the malignant or benign nature of an adnexal mass compared with all the other models considered.

Two recent multicentre validation studies (Timmerman et al., 2010; Van Holsbeke et al., 2012; Sayasneh et al., 2013b) compared the diagnostic accuracy of the best performing models according to this study (LR2 and SR) with the most frequently validated and used model RMI-1. A meta-analysis combining individual centre-specific results stratified for menopausal status from both multicentre cohorts showed a pooled sensitivity and specificity in premenopausal women for LR2 of 0.85 [95% CI 0.75–0.91] and 0.91 [95% CI 0.83–0.96] compared with 0.93 [95% CI 0.84–0.97], 0.83 [95% CI 0.73–0.90] for SR and 0.44 [95% CI 0.28–0.62], 0.95 [95% CI 0.90–0.97] for RMI-1. In post-menopausal women, the observed differences were smaller. The sensitivity and specificity of LR2, SR and RMI-1 were 0.94 [95% CI 0.89–0.97] and 0.70 [95% CI 0.62–0.77], 0.93 [95% CI 0.88–0.96] and 0.76 [95% CI 0.69–0.82], and 0.79 [95% CI 0.72–0.85] and 0.90 [95% CI 0.84–0.94], respectively (Fig. 5A–F).

The performance of the models that were excluded from the quantitative data synthesis and their references (i.e. models that were only described in their original publication without further validation, models that were only validated once, or in a limited number of adnexal masses) can be found as supporting information to this review (Supplementary data, Tables SI and SII, and Text S2).

Discussion

Summary of evidence
Adding new studies from our extended search leads to different conclusions to those published previously (Geomini et al., 2009). Following our systematic review of the literature and according to the results of the meta-analysis, we suggest that the evidence-based approach to the preoperative characterization of adnexal pathology should include the use of an ultrasound-based prediction model such as the IOTA LR2 model or ultrasound-based SR.

Strengths and limitations
The strengths of our report include the incorporation of evidence that reflects the recent development of new diagnostic tests for the preoperative identification of malignant disease in women with adnexal masses, the strict adherence to recent guidelines for diagnostic reviews (Irwig et al., 1994; Deeks, 2001; Moher et al., 2009; Macaskill et al., 2010), and the use of advanced statistical methods to report summary estimates of sensitivity and specificity and summary ROC curves for the models under consideration (Macaskill et al., 2010).

Like any other systematic review, this study is not without limitations. First, in an ideal setting, comparative accuracy between all considered tests or models in this review should be estimated by restricting analysis to those studies that have evaluated these tests/models in the same individuals (Macaskill et al., 2010). However, there were too few studies that evaluated multiple tests in our meta-analysis to adopt such a general approach.

On the other hand, three recent validation studies did compare overall test performance of the IOTA LR2 model (Van Holsbeke et al., 2012; Sayasneh et al., 2013b) and IOTA SR (Sayasneh et al., 2013b) with both RMI-1 and the ROMA algorithm (Kaijser et al., 2013c) on the same set of patients. In the first study, a total of 742 (74%) benign and 255 (26%) malignant masses were included (Van Holsbeke et al., 2012). Showing an overall sensitivity of 92% and specificity of 86% LR2 achieved a higher diagnostic accuracy compared with the current reference test RMI-1 (sensitivity 67% and specificity 95%). The difference in AUC between LR2 and RMI-1 was 0.038 (95% CI 0.018–0.058). In
Table I  Prediction models (n = 19) considered in the quantitative data synthesis with modelling type, variables and the number of validation studies.

<table>
<thead>
<tr>
<th>Model</th>
<th>Type of model</th>
<th>Variables</th>
<th>Cut-off level</th>
<th>Number of validation studies</th>
<th>Number of adnexal masses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tailor</td>
<td>Logistic</td>
<td>(i) Papillations, (ii) age (iii) time-averaged maximum velocity in tumour vessels</td>
<td>50%</td>
<td>6</td>
<td>2527</td>
</tr>
<tr>
<td>LRA</td>
<td>Logistic</td>
<td>(i) Colour score, (ii) CA125, (iii) papillations (iv) menopausal score</td>
<td>25%</td>
<td>3</td>
<td>1888</td>
</tr>
<tr>
<td>LRB</td>
<td>Logistic</td>
<td>(i) Papillations, (ii) internal wall, (iii) unilocular cyst, (iv) ascites, (v) bilaterality, (vi) menopausal score, (vii) CA125</td>
<td>60%</td>
<td>5</td>
<td>2155</td>
</tr>
<tr>
<td>Jokubiene</td>
<td>Logistic</td>
<td>(i) Size of lesion (mean of 3 diameters), (ii) size of largest solid component (mean of 3 diameters), (iii) any irregularity</td>
<td>12%</td>
<td>2</td>
<td>2043</td>
</tr>
<tr>
<td>Prompeler</td>
<td>Logistic</td>
<td>(i) Ascites, (ii) solid lesion without shadowing, (iii) cyst with &gt;30% solid part, (iv) diameter of the lesion, (v) multilocularity, (vi) surface of the cyst</td>
<td>10%</td>
<td>2</td>
<td>1236</td>
</tr>
<tr>
<td>IOTA LR2</td>
<td>Logistic</td>
<td>(i) Age, (ii) ascites, (iii) blood flow within a papillary projection, (iv) maximal diameter of the largest solid component (bounded at 50 mm), (v) irregular internal cyst wall, (vi) acoustic shadows</td>
<td>10%</td>
<td>3</td>
<td>1356</td>
</tr>
<tr>
<td>Ferrazzi</td>
<td>Morphologic</td>
<td>(i) Wall structure, (ii) septa, (iii) vegetation, (iv) echogenicity</td>
<td>9</td>
<td>9</td>
<td>1814</td>
</tr>
<tr>
<td>Depriest</td>
<td>Morphologic</td>
<td>(i) Tumour volume, (ii) wall structure, (iii) septal structure</td>
<td>5</td>
<td>10</td>
<td>1957</td>
</tr>
<tr>
<td>Lerner</td>
<td>Morphologic</td>
<td>(i) Wall structure, (ii) acoustic shadows, (iii) septa, (iv) echogenicity</td>
<td>3</td>
<td>10</td>
<td>3035</td>
</tr>
<tr>
<td>Sassone</td>
<td>Morphologic</td>
<td>(i) Inner wall structure, (ii) wall thickness, (iii) septa, (iv) echogenicity</td>
<td>9</td>
<td>21</td>
<td>2981</td>
</tr>
<tr>
<td>RMI I</td>
<td>Multimodal</td>
<td>(i) Menopausal status, (ii) CA125, (ii) multilocular cysts, (iv) solid areas, (v) metastases, (vi) ascites, (vii) bilaterality</td>
<td>200</td>
<td>35</td>
<td>9597</td>
</tr>
<tr>
<td>RMI II</td>
<td>Multimodal</td>
<td>Same as RMI I</td>
<td>200</td>
<td>18</td>
<td>4772</td>
</tr>
<tr>
<td>RMI III</td>
<td>Multimodal</td>
<td>Same as RMI I</td>
<td>200</td>
<td>14</td>
<td>5169</td>
</tr>
<tr>
<td>RMI IV</td>
<td>Multimodal</td>
<td>(i) Menopausal status, (ii) CA125, (ii) multilocular cysts, (iv) solid areas, (v) metastases, (vi) ascites, (vii) bilaterality, (viii) largest diameter of lesion</td>
<td>450</td>
<td>5</td>
<td>2191</td>
</tr>
<tr>
<td>ANN1</td>
<td>Artificial neural network</td>
<td>(i) Papillations, (ii) colour score, (iii) menopausal status, (iv) CA125</td>
<td>45%</td>
<td>3</td>
<td>1976</td>
</tr>
<tr>
<td>ANN2</td>
<td>Artificial neural network</td>
<td>(i) Papillations, (ii) smooth surface, (iii) unilocular, (iv) ascites, (v) bilaterality, (vi) menopausal status, (vii) CA125</td>
<td>60%</td>
<td>4</td>
<td>2055</td>
</tr>
<tr>
<td>Simple Rules</td>
<td>Ultrasound rules</td>
<td>M criteria: (irregular solid mass, colour score 4, irregular multilocular-solid mass ≥ 100 mm, ascites, at least 4 papillary structures) B-criteria: (unilocular cyst, colour score 1, smooth multilocular tumour with largest diameter &lt; 100 mm, presence of acoustic shadows, presence of solid components where the largest solid component has a largest diameter &lt; 7 mm)</td>
<td>n/a</td>
<td>5</td>
<td>2315</td>
</tr>
<tr>
<td>ROMA</td>
<td>Biomarker</td>
<td>(i) CA125, (ii) HE4, (iii) menopausal status</td>
<td>n/a</td>
<td>18</td>
<td>5116</td>
</tr>
<tr>
<td>OVA-1*</td>
<td>Biomarker</td>
<td>(i) CA125-II, (ii) transferrin, (iii) transthyretin (prealbumin), (iv) apolipoprotein A1, (vii) beta-2-microglobulin</td>
<td>n/a</td>
<td>2</td>
<td>1018</td>
</tr>
</tbody>
</table>

LRA, logistic regression model a; LRB, logistic regression model b; IOTA LR2, logistic regression model 2 of the International Ovarian Tumour Analysis study; RMI, risk of malignancy index; ANN, artificial neural network; ROMA, risk of ovarian malignancy algorithm.

*Original publication of model.

†Included in quantitative data synthesis.

‡For simple rules, the inconclusive cases are considered as malignant.

§Cut-off value is different for pre and post-menopausal women.
the second study, which included 255 adnexal masses (malignancy rate 29%), the difference in AUC was 0.04 (95% CI 0.01–0.07) in favour of LR2 (Sayasneh et al., 2013b). In this same study, a strategy of IOTA SR with classifying all inconclusive masses as malignant had an overall sensitivity of 91% and specificity of 87% compared with 72 and 94% for RMI-1, respectively (Sayasneh et al., 2013b). In the third study a total of 144 (40%) malignant and 216 (60%) benign adnexal masses were included (Kaijser et al., 2013c). The overall test performance of LR2 (AUC 0.952) with a 94% sensitivity and 82% specificity was better than ROMA (AUC 0.893) with 84% sensitivity and 80% specificity. The difference in AUC was 0.059 (95% CI 0.026–0.091).

The second limitation was that this meta-analysis focused on sensitivity and specificity to define the diagnostic accuracy of a test. We acknowledge that it is not always correct to base universal recommendations for

<table>
<thead>
<tr>
<th>Model</th>
<th>Cut-off</th>
<th>Studies (n)</th>
<th>Centres* (n)</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphologic scoring systems</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sassone</td>
<td>&gt; 9</td>
<td>19</td>
<td>19</td>
<td>0.85 [0.77;0.90]</td>
<td>0.80 [0.73;0.86]</td>
</tr>
<tr>
<td>Lerner</td>
<td>&gt; 3</td>
<td>9</td>
<td>17</td>
<td>0.80 [0.70;0.86]</td>
<td>0.61 [0.53;0.68]</td>
</tr>
<tr>
<td>Depriest</td>
<td>&gt; 5</td>
<td>8</td>
<td>8</td>
<td>0.90 [0.81;0.95]</td>
<td>0.68 [0.57;0.77]</td>
</tr>
<tr>
<td>Ferrazzi</td>
<td>&gt; 9</td>
<td>7</td>
<td>7</td>
<td>0.86 [0.77;0.91]</td>
<td>0.80 [0.66;0.89]</td>
</tr>
<tr>
<td>Ultrasound rules</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Rules</td>
<td>n/a</td>
<td>5</td>
<td>17</td>
<td>0.93 [0.89;0.95]</td>
<td>0.81 [0.76;0.85]</td>
</tr>
<tr>
<td>Risk of Malignancy Indexes (RMI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMI I</td>
<td>200</td>
<td>23</td>
<td>41</td>
<td>0.72 [0.67;0.76]</td>
<td>0.92 [0.89;0.93]</td>
</tr>
<tr>
<td>RMI II</td>
<td>200</td>
<td>15</td>
<td>32</td>
<td>0.75 [0.69;0.80]</td>
<td>0.87 [0.84;0.90]</td>
</tr>
<tr>
<td>RMI III</td>
<td>200</td>
<td>9</td>
<td>19</td>
<td>0.70 [0.60;0.78]</td>
<td>0.91 [0.88;0.93]</td>
</tr>
<tr>
<td>RMI IV</td>
<td>450</td>
<td>3</td>
<td>13</td>
<td>0.68 [0.59;0.76]</td>
<td>0.94 [0.91;0.96]</td>
</tr>
<tr>
<td>Logistic regression models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tailor</td>
<td>50%</td>
<td>6</td>
<td>24</td>
<td>0.35 [0.24;0.49]</td>
<td>0.96 [0.94;0.98]</td>
</tr>
<tr>
<td>LRa</td>
<td>25%</td>
<td>3</td>
<td>20</td>
<td>0.76 [0.70;0.81]</td>
<td>0.87 [0.82;0.90]</td>
</tr>
<tr>
<td>LRb</td>
<td>60%</td>
<td>4</td>
<td>21</td>
<td>0.82 [0.77;0.86]</td>
<td>0.78 [0.73;0.83]</td>
</tr>
<tr>
<td>Prömpeler</td>
<td>10%</td>
<td>2</td>
<td>10</td>
<td>0.61 [0.46;0.74]</td>
<td>0.81 [0.70;0.89]</td>
</tr>
<tr>
<td>Jokubkiene</td>
<td>12%</td>
<td>2</td>
<td>20</td>
<td>0.77 [0.71;0.82]</td>
<td>0.87 [0.83;0.89]</td>
</tr>
<tr>
<td>IOTA LR2</td>
<td>10%</td>
<td>3</td>
<td>13</td>
<td>0.92 [0.88;0.95]</td>
<td>0.83 [0.77;0.88]</td>
</tr>
<tr>
<td>Artificial neural networks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN1</td>
<td>45%</td>
<td>3</td>
<td>20</td>
<td>0.77 [0.71;0.82]</td>
<td>0.86 [0.80;0.90]</td>
</tr>
<tr>
<td>ANN2</td>
<td>60%</td>
<td>4</td>
<td>21</td>
<td>0.97 [0.95;0.98]</td>
<td>0.37 [0.31;0.44]</td>
</tr>
</tbody>
</table>

*For the centre-specific results, only centres of multicentre studies that contributed at least 3 benign and 3 malignant cases are considered.

bFor simple rules, the inconclusive cases are considered as malignant.

Figure 3 Forest plot demonstrating validation studies of the ROMA algorithm. Test performance of ROMA when differentiating between all benign and malignant disease. Sens, sensitivity; LCL, lower confidence limit; UCL, upper confidence limit; Spec, specificity.
examiners replicate the results of ‘experts’. There is now evidence that the choice of a clinical test upon these criteria. The adopted cut-offs vary between models, and we recognize that this complicates interpretation of the results as the cut-off entails an assumed relative preference for TP versus TN results (Van Calster et al., 2013).

Third, we have been unable to investigate the effect of possible sources of heterogeneity with meta-regression due to the limited number of validation studies for several prediction models. The pooled estimates of overall diagnostic test performance of our considered models should therefore be interpreted with a degree of caution. The quality assessment in our report did reveal shortcomings, especially in study design (e.g. retrospective data collection, absence of adequate verification) and quality of reporting (e.g. blinding issues, delay between index and reference test). In total 41% of all included studies were retrospective in design or did not properly report this quality item. A lack of adequate blinding (e.g. knowledge of the reference test) in these retrospective studies can bias relatively subjective index tests like transvaginal ultrasonography, and lead to overoptimistic results of diagnostic test performance (Lijmer et al., 1999). On the other hand, the IOTA prediction models have been subject to external validation on a large sample of patients in studies with prospective and consecutive data sampling and adequate blinding procedures (Timmerman et al., 2010; Fathallah et al., 2011; Nunes et al., 2012; Van Holsbeke et al., 2012; Alcazar et al., 2013; Sayasneh et al., 2013b).

A general criticism of the use of transvaginal ultrasonography as a diagnostic test for any woman with possible ovarian cancer is that it is subjective and its performance is dependent upon the experience and skills of the operator. The main aim of the IOTA study was to tackle this problem and produce models and rules that may help less experienced examiners replicate the results of ‘experts’. There is now evidence available that suggests that ultrasound-based prediction models and rules retain their performance in the hands of both relatively inexperienced doctors and sonographers (Nunes et al., 2012; Alcazar et al., 2013; Sayasneh et al., 2013a, b). Therefore, it should be possible for most ultrasound examiners to characterize the majority of ovarian masses with an acceptable level of test performance. In a recent multicentre validation study by operators with varied training, test performance for LR2 was better in premenopausal women than in post-menopausal women (diagnostic odds ratio of 121 and 23 in pre- and post-menopausal, respectively) (Sayasneh et al., 2013b). Similarly, when SR + SA (Simple Rules followed by subjective assessment of level II ultrasonography examiners when the rules were inconclusive) were applied, the diagnostic performance of this approach was better in premenopausal compared with post-menopausal women (diagnostic odds ratio of 100 and 82, respectively) (Sayasneh et al., 2013b). These findings are of importance, as this validation study suggests that the IOTA models and rules (LR2 and SR) retain their performance in premenopausal women in the hands of less experienced (level II) examiners. In contrast RMI-1 had only a diagnostic odds ratio of 30 and 37 in pre- and post-menopausal women, respectively (Sayasneh et al., 2013b). The discriminative performance of the IOTA models in women of reproductive age should optimize fertility preservation in these women.

Unanswered questions and future research

Almost all diagnostic tests considered in this review, including the IOTA LR2 model and Simple rules, are based on women scheduled for surgery and use final histology as their reference standard. No information is available as to what percentages of women with masses were triaged for expectant management and if managed conservatively what the long-term outcome was for this group of women with respect to malignant transformation and possible complications such as torsion, rupture and haemorrhage which might also jeopardize ovarian function (Alcazar et al., 2005). Accordingly there is a need for studies that examine the long-term outcome for masses classified as benign and treated expectantly without surgical intervention. In addition, more research is needed to develop risk prediction tools for multiple outcomes such as BOTs, early versus advanced stage invasive ovarian cancer and non-ovarian cancer metastasized to the ovary. Most prediction models used in clinical practice use a dichotomous outcome (i.e. benign or malignant). However, various subtypes of malignant disease are managed differently with implications for the type of surgery (i.e. fertility sparing
**Figure 5** Subgroup analysis of test performance of LR2, Simple Rules and RMI-1 in premenopausal (A, B and C, respectively) and post-menopausal women (D, E and F, respectively). LR2, logistic regression model 2; RMI-1, risk of malignancy index 1; SR, simple rules; Sens, sensitivity; LCL, lower confidence limit; UCL, upper confidence limit; Spec, specificity; GBE, Genk, Belgium; BCH, Beijing, China; PCR, Prague, Czech Republic; BIT, Bologna, Italy; CIT, Milano, Italy; VIT, Milano, Italy; GIT, Napoli, Italy; SIT, Sardinia, Italy; UDI, Udinese, Italy; LPO, Lublin, Poland; LSW, Lund, Sweden; QCCH, Queen Charlotte’s and Chelsea Hospital, London, UK; SHH, Princess Anne Hospital, Southampton, UK.
surgery for BOT and early stage invasive ovarian cancer versus radical surgery), length of hospitalization and financial cost (Kaiser et al., 2013a).

**Conclusion**

This update of a previous systematic review and meta-analysis shows that both IOTA LR2 and Simple Rules are currently the best diagnostic tests available for differentiating between the benign or malignant nature of an adnexal mass in a preoperative setting. In women of reproductive age, such differentiation is pivotal for preservation of fertility. Accordingly, healthcare policy makers should review current triaging protocols.

**Supplementary data**

Supplementary data are available at http://humupd.oxfordjournals.org/.

**Authors’ roles**

D.T., T.B. and S.G.-M. developed the idea for this study, J.K. and A.S. performed the systematic search, extraction of data and quality assessment, and K.V.H. and B.V.C. carried out the statistical analysis. J.K., A.S., K.V.H., B.V.C., S.G.-M., T.B. and D.T. all contributed to drafting and revising the manuscript critically for intellectual content. All authors approved the final version of the manuscript. All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis. D.T., T.B. and B.V.C. act as final guarantors of the paper.

**Funding**

K.V.H. is supported by a PhD grant of the Flanders’ Agency for Innovation by Science and Technology (IWT Vlaanderen). B.V.C. is a postdoctoral fellow from the Research Foundation, Flanders (FWO). D.T. and B.V.C. received a project grant from the FWO (grant G049312N) for the IOTA 5 project. T.B. and S.G.M. are supported by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Conflict of interest**

All authors declare no support from any commercial organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

**References**


Anton C, Carvalho FM, Oliveira El, Maciel GA, Barata EC, Carvalho JP. A comparison of CA 125, HE4, risk ovarian malignancy algorithm (ROMA), and risk malignancy index (RMI) for the classification of ovarian masses. Clinics (Sao Paulo) 2012; 67:437–441.


Davies AP, Jacobs I, Woolas R, Fish A, Oram D. The adnexal mass: benign or malignant?


