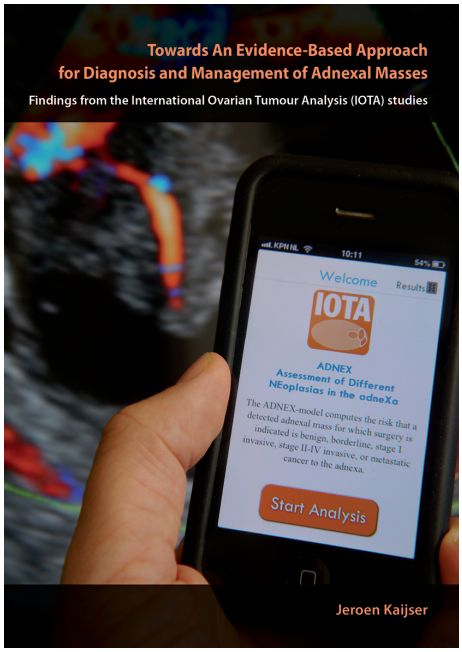
Towards an evidence-based approach for diagnosis and management of adnexal masses: findings of the International Ovarian Tumour Analysis (IOTA) studies

J. Kaijser[1,2]
Promotor:T. Bourne[1,2,3]
Co-promotors: B. Van Calster[1], D. Timmerman[1,2]



[1]KU Leuven, Department of Development and Regeneration, Leuven, Belgium; [2]Department of Obstetrics and Gynaecology, University Hospitals KU Leuven, Leuven, Belgium; [3]Queen Charlotte's & Chelsea Hospital, Imperial College, Du Cane Road, London W12 0HS, United Kingdom.

Correspondence at: Jeroen Kaijser, Department of Obstetrics and Gynecology, University Hospitals KU Leuven, Herestraat 49, 3000 Leuven, Belgium. E-mail: jeroen.kaijser@uzleuven.be

Abstract

Whilst the outcomes for patients with ovarian cancer clearly benefit from centralised, comprehensive care in dedicated cancer centres, unfortunately the majority of patients still do not receive appropriate specialist treatment. Any improvement in the accuracy of current triaging and referral pathways whether using new imaging tests or biomarkers would therefore be of value in order to optimise the appropriate selection of patients for such care. An analysis of current evidence shows that such tests are now available, but still await recognition, acceptance and widespread adoption. It is therefore to be hoped that present guidance relating to the classification of ovarian masses will soon become more "evidence-based". These promising tests include the International Ovarian Tumour Analysis (IOTA) LR2 model and ultrasound-based Simple Rules (SR). Based on a comprehensive recent meta-analysis both currently offer the optimal "evidence-based" approach to discriminating between cancer and benign conditions in women with adnexal tumours needing surgery. LR2 and SR are reliable tests having been shown to maintain a high sensitivity for cancer after independent external and temporal validation by the IOTA group in the hands of examiners with various levels of ultrasound expertise. They also offer more accurate triage compared to the existing Risk of Malignancy Index (RMI). The development of the IOTA ADNEX model represents an important step

forward towards more individualised patient care in this area. ADNEX is a novel test that enables the more specific subtyping of adnexal cancers (i.e. borderline, stage 1 invasive, stage II-IV invasive, and secondary metastatic malignant tumours) and shares similar levels of accuracy to IOTA LR2 and SR for basic discrimination between cancer and benign disease. The IOTA study has made significant progress in relation to the classification of adnexal masses, however what is now needed is to see if these or new diagnostic tools can assist clinicians to select patients with adnexal masses that are suitable for expectant management, and that will work in all health care settings (i.e. primary vs secondary vs tertiary care). These important themes will likely control the future agenda of the IOTA project.

Key words: **Ovarian cancer, diagnostic test, triage, classification, systematic review, meta-analysis.**

Introduction

The International Ovarian Tumour Analysis (IOTA) study is the largest diagnostic accuracy study of its kind. The group running the study set out more than 15 years ago with the ultimate goal of developing an optimal "evidence-based" algorithm for the classification and management of all types of adnexal masses. Such an algorithm should enable clinicians to direct treatment for those patients with ovarian cancer to subspecialist gynaecological oncologists whilst facilitating the selection of patients with benign conditions for either ultrasound follow-up or minimal access surgery if there is a need to intervene. A total of 47 centres from 17 countries (Fig. 1) have now contributed to one or more of its different phases recruiting over 10.000 patients with even a larger number of adnexal masses, and the consortium is likely to expand to include greater numbers in the near future.

Before the start of this thesis, IOTA has already brought many benefits to the field of transvaginal ultrasonography (TVS) and classification of adnexal tumours, not least a standardised approach with clear-cut definitions and qualitative and quantitative end-points to describe both the morphological and Doppler ultrasound features of these tumours (Campbell, 2012; Timmerman et al., 2000). This agreement on a unified approach to describe all possible ultrasound variables is a key element to future successful implementation of any ultrasound based protocol for ovarian pathology (Kaijser et al., 2012). The IOTA ultrasound protocol and the subsequent large amount of information collected prospectively (2000-2007) about patients and tumours included in IOTA phase 1 (n = 1066), phase 1a (n = 507) and phase 2 (n = 1938) enabled existing prediction models for ovarian cancer like the Risk of Malignancy Index (RMI) (Jacobs et al., 1990) to be tested, and novel promising ultrasound-based strategies (risk models (IOTA LR1, LR2) and Simple Rules) to be developed in IOTA phase 1 and externally validated in IOTA phase 2 by expert examiners (Timmerman et al., 2005, 2008; Van Holsbeke et al., 2009; Timmerman et al., 2010a; Timmerman et al., 2010b; Van Holsbeke et al., 2012). These approaches when used by expert

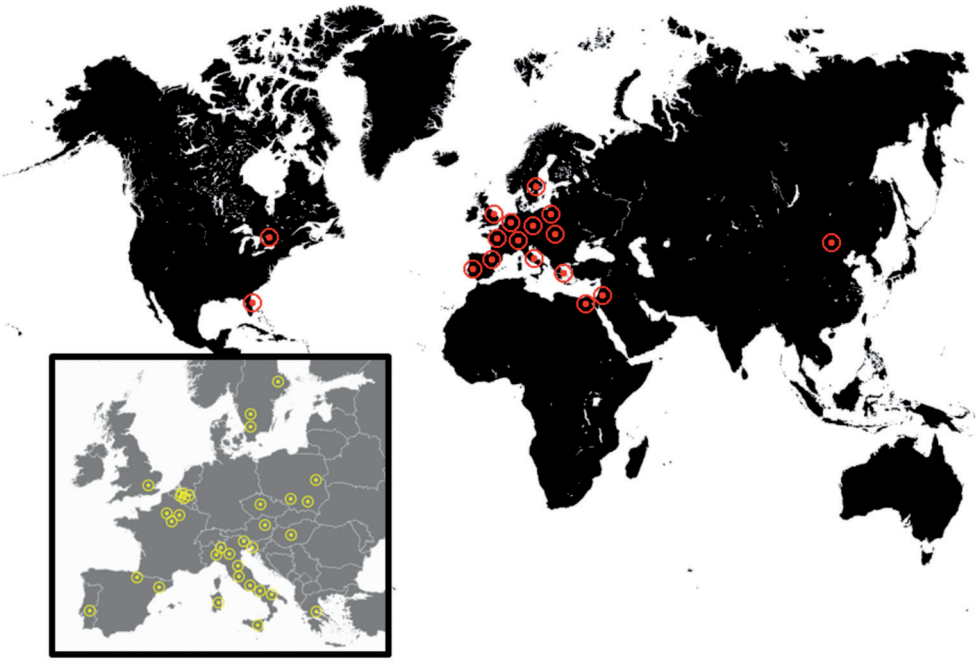


Fig. 1. — The International Ovarian Tumour Analyis (IOTA) study network of participating centres

sonographers have shown to come close to matching test performance of their own pattern recognition or subjective assessment of ultrasound and colour Doppler findings. Both risk models also showed significantly better test performance than the RMI in IOTA phase 2 (Van Holsbeke et al., 2012). In addition, the results of IOTA phase 1 and 2 studies have also showed that a single measurement of serum CA-125 using a fixed cut-off (35 IU/l) has a rather limited role to play in the preoperative discrimination between cancer and benign ovarian pathology, either when used as a single test (Van Calster et al., 2007), incorporated in logistic regression models (Timmerman et al., 2007), or as a secondary test in masses found difficult to classify by expert ultrasound examiners (Valentin et al., 2009).

Whether the main IOTA strategies (IOTA LR2 and SR) will stop the diagnostic algorithm mania (Campbell, 2012) that has dominated approaches to the preoperative classification of adnexal tumours and become the standard of care to triage women to appropriate patientcare pathways had not been definitively answered yet before the start of this thesis. Endorsement by being included in clinical guidelines, acceptance and widespread adoption would require:

1) Independent validation by other research groups.
2) Preferably validation by less experienced operators (IOTA phase 4: 2010-2013).
3) A critical appraisal and summary of these studies in systematic reviews by means of meta-analysis.
4) Direct comparison to other existing approaches such as the RMI (Jacobs et al., 1990). (IOTA phase 3: 2009-2012) and more recent emerging imaging or biomarker algorithms (e.g. the Risk of Ovarian Malignancy Algorithm (ROMA)) (Moore et al., 2009).

Besides classical preoperative classification of adnexal masses (benign versus malignant) there were still other important related issues to be solved by the IOTA group.

1) Predicting whether a mass is benign or malignant is not the only clinically relevant endpoint that we need to know before deciding on appropriate treatment. Besides classical factors like age, comorbidity, and the presence of symptoms, knowing the specific histopathology (i.e. dermoid cyst, endometrioma, mucinous borderline tumour of the intestinal type, stage 1 invasive serous cancer, Kruckenberg metastasis) of a mass is becoming of increasing importance to tailor management and treatment options to the individual patient. Preliminary research by the IOTA group already demonstrated feasibility of such a polytomous approach to the classification of adnexal tumours, but it failed to discriminate between two very important groups: invasive cancer confined to the ovary (Stage I) and advanced stage disease (Stage II-IV) (Van Calster et al., 2010).
2) Are other imaging techniques like magnetic resonance imaging (MRI) or new biomarkers like serum human epididymis protein-4 (HE4) or the ROMA algorithm helpful to classify a subgroup of difficult to classify adnexal tumours? At present around 8% of tumours are difficult to diagnose for expert examiners and would definitively benefit from an accurate secondary test (Valentin et al., 2006). However, so far none proved helpful to reliably diagnose these tumours (Valentin et al., 2006, 2011).
3) Do we need to treat every tumour classified as being benign either by using models or our subjective assessment of ultrasound findings? The majority of these masses are asymptomatic and probably do not pose any threat. However whilst the answer seems straightforward, the natural history of cysts that do not harbour any features of malignancy is still largely unknown and a relatively unexplored research area. There is an urgent need for robust evidence on the long-term outcome of cysts that are not thought to require surgery – both in terms of long term risk of malignancy, but also the incidence of cyst accidents (IOTA phase 5: 2012 - present). This crucial information should enable the IOTA study to develop a single test that can be used in different healthcare settings and will improve clinical decision-making in patients with all types of adnexal tumours.

Objective of the Thesis

The main objective of this doctoral thesis was to describe the main findings of 15 years of clinical research by the International Ovarian Tumour Analysis study group. The results of IOTA phases 1, 1b and 2 (2000-2007) have already been summarised above in the introduction (Kaijser et al., 2013a). The objective of this thesis was further broken down into more specific aims:

1.1 The issue of operator experience and (independent) validation of IOTA strategies (IOTA phase 3 and 4: 2009 - 2013).

1.2 A critical appraisal of evidence on diagnostic tests for preoperative diagnosis of ovarian cancer: a systematic review and meta-analysis.

1.3 Multiclass risk prediction of adnexal tumours (IOTA phase 1-3: 2000-2012).

Table I. — Test performance of logistic regression model 2 (LR2) and Simple Rules derived by the International Ovarian Tumour Analysis (IOTA) group for discrimination between benign and malignant adnexal masses and of the risk of malignancy index (RMI) in each phase of the IOTA study.

Models	IOTA phase	Type of validation	Experience level of ultrasound examiner	N	Sens (%)	Spec (%)	LR+	LR-	DOR	AUC
LR2 (10% cut off)	1	Development data	Expert	754	92	75	3.71	0.10	35.5	0.93
	1	Internal (test set)	Expert	312	89	73	3.36	0.15	23.1	0.92
	1b	Temporal	Expert	507	95	74	3.64	0.07	55.0	0.95
	2	Temporal	Expert	941	89	80	4.42	0.14	32.7	0.92
	2	External	Expert	997	92	86	6.36	0.10	66.1	0.95
	3	Temporal	Expert	2403	91	78	4.13	0.11	36.6	0.93
	4	External	Non-expert	255	88	90	8.37	0.14	61.6	0.94
Simple Rules with subjective assessment[a]	1	Development data	Expert	1066	91	90	8.84	0.10	84.4	N/A
	1b	Temporal	Expert	507	92	90	9.08	0.09	106	N/A
	2	Temporal	Expert	941	92	93	12.28	0.09	142	N/A
	2	External	Expert	997	90	93	12.63	0.11	120	N/A
	3	Temporal	Expert	2403	93	88	7.77	0.09	90.8	N/A
	4	External	Non-expert	255	86	94	15.65	0.14	109.4	N/A
Simple Rules (classifying inconclusive cases as malignant)	1	Development data	Expert	1066	95	65	2.71	0.08	33.4	N/A
	1b	Temporal	Expert	507	96	68	3.03	0.06	49.4	N/A
	2	Temporal	Expert	941	95	74	3.74	0.06	61.5	N/A
	2	External	Expert	997	94	80	4.76	0.08	61.0	N/A
	3	Temporal	Expert	2403	95	73	3.58	0.06	55.9	N/A
	4	External	Non-expert	255	91	87	7.13	0.11	65.8	N/A
Risk of Malignancy Index (200 cut off)	2	External	Expert	997	67	95	12.7	0.34	36.8	0.91
	3	External	Expert	2403	70	90	7.11	0.33	21.7	0.90
	4	External	Non-expert	255	72	94	12.96	0.30	43.2	0.90

LR 2: logistic regression model 2; AUC: area under the receiver-operator characteristics curve (ROC); DOR: diagnostic odds ratio; LR+: positive likelihood ratio; LR-: negative likelihood ratio. RMI: risk of malignancy index. N/A: not applicable.
[a] Results are shown for simple rules supplemented with subjective assessment of ultrasound findings when the rules did not apply.

1.4 Serum HE4 and the ROMA algorithm for ovarian cancer diagnosis.
1.5 Imaging techniques to diagnose adnexal tumours.

Results and Discussion of main findings

1.1 The issue of operator experience and validation of IOTA strategies

Within the framework of the IOTA study the main ultrasound-based approaches (LR2 and SR) had been successfully externally validated by level III expert ultrasound examiners (according to European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) guidelines (EFSUMB, 2006) in IOTA phase 2 on a total of 997 adnexal masses collected between 2005 and 2007 (Timmerman et al., 2010b; Van Holsbeke et al., 2012). Both strategies retained their excellent test performance regarding discrimination of adnexal masses and significantly outperformed the RMI. A second large temporal validation by those same level III expert examiners was conducted on 2403 new cases collected between 2009 and 2012 in IOTA phase 3 and confirmed the previous findings of test accuracy in IOTA phase 2 (Table I) (Testa et al., 2014).

The use of a meta-analysis technique on the IOTA phase 3 multicentre dataset also enabled us to demonstrate centre-specific differences in test performance (AUC, sensitivity, specificity) and calibration results for LR2, Simple Rules, and RMI (Testa et al., 2014). In previous phases of IOTA, results for diagnostic accuracy were reported as single pooled estimates (Timmerman et al., 2005,

2008, 2010a, 2010b; Van Holsbeke et al., 2009, 2012). In IOTA phase 3 a total of 7 tertiary oncology referral centres and 11 local community hospitals were included with marked differences in cancer prevalence varying from 0 to 69% (Testa et al., 2014). Both IOTA methods (LR2 and SR) had a higher sensitivity for cancer than RMI, irrespective of the prevalence of malignancy (Testa et al., 2014). Our use of meta-analysis techniques to summarise data did not meaningfully change the summary measures of test performance from those obtained with a standard pooled analysis but gave wider confidence intervals properly reflecting the uncertainty caused by differences between centres (Testa et al., 2014). The IOTA LR2 model showed consistently high AUCs (discrimination) across different types of centres, but tended to underestimate the true risk of cancer (calibration), especially in those centres with a high prevalence of cancer (Testa et al., 2014). Ideally, a risk model should have perfect discrimination and calibration. When interpreting risks (IOTA LR2) for individual patient management within the context of a single centre some caution is thus needed. The most plausible explanations for these findings are differences in tumour mix, equipment, and examiners' use of the IOTA terms (Testa et al., 2014). This has recently been demonstrated in a real-time inter-observer study by Sladkevicius and Valentin (2014). In this study the inter-observer agreement in classifying tumours as benign or malignant (discrimination) using the risk of malignancy cut off of 10% for LR1 and LR2 was good. However, risk estimates differed substantially (calibration) between sonologists (Sladkevicius and Valentin, 2014). But this seems inherent to all prediction models that incorporate rather subjective ultrasound criteria, for instance the RMI. RMI in turn does not provide clinicians with risks but "calibration" using IOTA 3 data did reveal that the implicit average risk of cancer at an RMI value of 200 (cut off to indicate malignancy) is around 50% (Testa et al., 2014). This implies that patients with a cancer risk of up to 50% may not be referred to gynaecological oncologists. Even for objective biomarkers like serum CA-125 the observed proportion of malignancies at the commonly used cut-off level of 35 IU/L varies from 10 up to 60 % depending on the type of centre in which the biomarker is used (unpublished findings).

A general criticism of the use of TVS as a diagnostic test for 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 robust and reliable models and rules that may help less experienced examiners replicate the results of "experts" (Kaijser et al., 2013a). In IOTA phase 4, a total of 35 "level II" ultrasound examiners (62.9% of the operators had performed <1000 ultrasound scans; 24% of the operators were medical doctors, whereas 76% were sonographers) prospectively validated SR, the IOTA LR2 model, and the IOTA three-step strategy (Simple Descriptors as a first stage test, SR for tumours in which the Simple Descriptors are not applicable, and subjective assessment for masses in which Simple Rules are inconclusive) (Ameye et al., 2012) on a total of 255 and 301 new cases, respectively, collected between 2010 and 2012 from three hospitals in the United Kingdom (UK) (Sayasneh et al., 2013a, 2013b). Both IOTA tests (SR and IOTA LR2) had excellent diagnostic accuracy, and even showed significantly better test performance than RMI in non-expert hands (Table I) (Sayasneh et al., 2013a). Findings were even comparable to performance of level III expert examiners. 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 IOTA 4 findings will be generalizable to most clinicians in daily practice (Sayasneh et al., 2013a). However, a weakness common to other studies is the difficulty encountered in classifying operator experience (Sayasneh et al., 2013a). The EFSUMB guidelines (EFSUMB, 2006) have tried to produce objective criteria to estimate the level of experience of ultrasound examiners, but these criteria do not apply to the 27 sonographers scanning patients in IOTA 4 (Sayasneh et al., 2013a). On the other hand level III operators (doctors) were excluded from IOTA 4. Whilst IOTA phase 4 has demonstrated that non-expert operators are perfectly capable of retrieving the information needed for the ultrasound variables required for IOTA LR2 or SR we should acknowledge that these reassuring findings were derived from units with good supervision, and an adequate knowledge of IOTA standardised terminology and definitions remains essential for any operator that would like to use LR2 or SR in clinical practice. All ultrasound examiners in IOTA 4 attended a half-day theoretical induction session where the ultrasound features of the rules and models used in the study were illustrated (Sayasneh et al., 2013a). As with many other clinical areas, the uncontrolled use of TVS and IOTA methods in clinical practice will lead to poor test performance and be likely to harm patients with adnexal tumours.

Other studies unrelated to the IOTA group have also demonstrated that the IOTA models and rules are reliable tests to triage patients in the hands of examiners with various backgrounds and levels

Table II. — Test performance of different Simple Rules (SR) strategies in the International Ovarian Tumour Analysis (IOTA) studies and after independent external validation.

External Validation Study	N	Prevalence of cancer in studied population	SR applicable (%)	Sens (%) if SR are applicable	Spec (%) if SR are applicable	Sens (%) SR+MA	Spec (%) SR+MA	Sens (%) SR+SA	Spec (%) SR+SA
Timmerman 2010	997	22	80	91	96	94	80	90	93
Fathallah 2011	122	12	89	73	97	79	88	n/a	n/a
Sayasneh 2013	255	29	84	87	98	91	87	86	94
Alcázar 2013	340	16	79	88	97	93	81	89	96
Hartman 2012	103	29	88	91	87	94	76	84	86
Nunes 2014	303	45	78	96	89	97	70	94	90
Tantipalakorn 2014	398	33	80	83	95	87	81	n/a	n/a

SR: simple rules; Sens: sensitivity; Spec: specificity; SR + MA: a strategy using simple rules and classifying those tumours where the rules yield an inconclusive result as malignant; SR + SA: a two-step strategy using simple rules as a triage test and subjective assessment for tumours in which the rules are inconclusive.

of experience. This especially applies to the IOTA Simple Rules. This is evidenced by the publication of seven external validation studies at 19 different clinical centres in various countries (including data from IOTA phase 2 and 4) (Timmerman et al., 2010b; Fathallah et al., 2011; Hartman et al., 2012; Alcázar et al., 2013; Sayasneh et al., 2013a; Nunes et al., 2014; Tantipalakorn et al., 2014). These studies have included a total of 2518 adnexal tumours (Timmerman et al., 2010b; Fathallah et al., 2011; Hartman et al., 2012; Alcázar et al., 2013; Sayasneh et al., 2013a; Nunes et al., 2014; Tantipalakorn et al., 2014). The proportion of tumours where SR could be applied varied from 79 to 89% (Timmerman et al., 2010b; Sayasneh et al., 2013a; Fathallah et al., 2011; Hartman et al., 2012; Alcázar et al., 2013; Nunes et al., 2014; Tantipalakorn et al., 2014). The reported sensitivity for cancer if the Rules applied varied from 73 to 96% with a specificity of 87 to 98% (Table II). In the studies of (Hartman et al., 2012; Alcázar et al., 2013; Sayasneh et al., 2013a; Nunes et al., 2014) the rules were applied by non-expert examiners. In their hands diagnostic accuracy was comparable to level III expert examiners (Timmerman et al., 2010b). The level of experience in the study of (Fathallah et al., 2011; Tantipalakorn et al., 2014) was not clearly stated.

If SR were used as a single test for all tumours based on a policy of classifying all inconclusive tumours where the rules do not apply as malignant, the reported sensitivity for cancer increases with a predictable cost of reduced specificity (Table II). Five out of these seven studies also assessed the test performance of a two-step approach using subjective assessment of both the B-mode ultrasound and colour Doppler findings as a secondary test to characterise those masses where the rules do not apply (Timmerman et al., 2010b; Hartman et al., 2012; Alcázar et al., 2013; Sayasneh et al., 2013a; Nunes et al., 2014). In this scenario the available data suggest that the false positive rate is reduced dramatically with only a slight decrease in the sensitivity for cancer (Table II). In three out of these five studies the rules were initially applied by non-experts, and the more difficult group of inconclusive tumours subsequently reviewed by experts (Hartman et al., 2012; Alcázar et al., 2013; Nunes et al., 2014). In the studies by (Sayasneh et al., 2013a; Timmerman et al., 2010b) both SR and subjective assessment were applied by the same examiners (level II and level III examiners, respectively).

The IOTA LR2 model seems less popular and attractive to clinicians, possibly because they do not feel the equation required to use the model has been easily available, something that has been addressed by the development of user friendly APPs. Although LR2 has only been externally validated in 3 studies (Van Holsbeke et al., 2012; Nunes et al., 2013; Sayasneh et al., 2013a), these include two multicentre IOTA studies (phase 2 and 4) (Van Holsbeke et al., 2012; Sayasneh et al., 2013a), in 13 different clinical environments on a total of 1544 adnexal masses. In two out of these three studies the test performance of LR2 was assessed in the hands of examiners with different backgrounds (doctors and sonographers) and level of expertise (Level II) (Nunes et al., 2013; Sayasneh et al., 2013a). The discriminative performance of LR2 (AUC of 0.93 and 0.94) was retained in their hands compared to results by experienced ultrasound examiners with a special interest in gynaecological ultrasonography (AUC of 0.95) (Van Holsbeke et al., 2012).

1.2 A critical appraisal of evidence on diagnostic tests for preoperative diagnosis of ovarian cancer: a systematic review and meta-analysis

As shown consistently in IOTA phases 1 to 3, the subjective assessment of ultrasound findings by experienced ultrasound examiners (level III) with a special interest in gynaecological ultrasonography is the best method to discriminate between malignant and benign adnexal tumours with a sensitivity and specificity of around 90% (Kaijser et al., 2013a; Testa et al., 2014). An alternative approach is to use diagnostic rules and models to triage women with adnexal tumours as being at low or high risk of cancer. This approach may be useful as ultrasound expertise is not always available in local community hospitals. These decision support tools have been developed to assist clinicians with variable training backgrounds and levels of expertise. In 2014 the IOTA group published the results of a systematic review and meta-analysis that aimed to determine the optimal test available at that time to characterise adnexal tumours prior to surgery (Kaijser et al., 2014a). Out of a total of 19 different prediction models, externally validated in 96 studies on 26 438 adnexal masses, including 7199 (27%) malignant and 19 239 (73%) benign masses, this meta-analysis concluded that the IOTA LR2 model and SR currently offer the optimal evidence-based approach to discriminate between malignant disease and benign conditions of the adnexae prior to surgery (Fig. 2) (Kaijser et al., 2014a).

Both IOTA models performed significantly better than the RMI (sensitivity of 72 % (95% CI 67-76%) and specificity of 92% (95% CI 89-93) using the 200 cut-off (Kaijser et al., 2014a), which currently is still recommended in many national guidelines by professional societies (RCOG, 2003; RCOG, 2011; NICE, 2011; SIGN, 2013). The IOTA LR2 model using a cut-off of 10% had a sensitivity of 92 % (95 % CI 88-95 %) and specificity of 83 % (95 % CI 77-88 %)(Kaijser et al., 2014a) when pooling the individual findings of three validation studies (Van Holsbeke et al., 2012; Sayasneh et al., 2013a; Nunes et al., 2012). Simple Rules had a pooled overall sensitivity of 93 % (95 % CI 89-95 %) and specificity of 81 % (95 % CI 76-85 %) for all masses when using a strategy that classifies inconclusive tumours as malignant and when individual results from the studies of (Timmerman et al., 2010b; Fathallah et al., 2011; Hartman et al., 2012; Alcázar et al., 2013; Sayasneh et al., 2013a) were combined. The findings of Nunes et al. (2014) and Tantipalakorn et al. (2014) were not included since these studies were not published when analysing our data. When results were stratified for menopausal status using

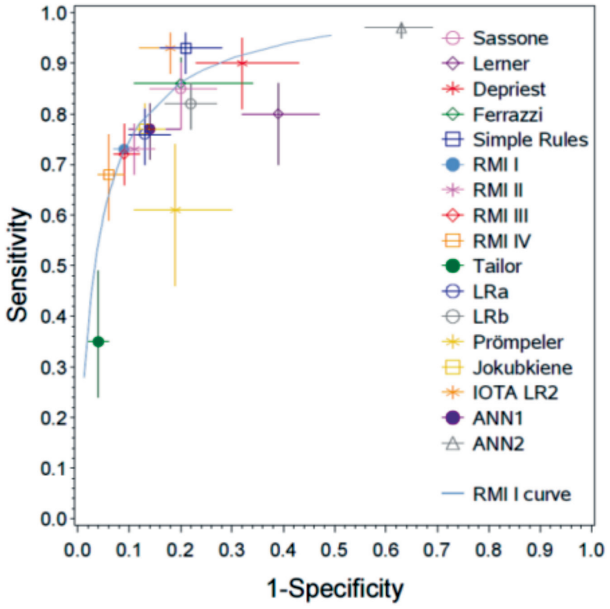


Fig. 2. — Overview of summary point estimates for the prediction models considered in the meta-analysis by the International Ovarian Tumour Analysis (IOTA) group.
All summary point estimates are displayed as sensitivity/specificity pairs with 95% CIs. For RMI-1 an additional hierarchical summary receiver operating curve (HSROC) was fitted. Reproduced with permission from *Kaijser J et al. Human Reprod Update 2014;20(3):449-62*.

multicentre data from IOTA phase 2 and 4 both IOTA LR2 and SR had a higher sensitivity for cancer than RMI in both pre and postmenopausal women, with a clear advantage in premenopausal women (Timmerman et al., 2010b; Van Holsbeke et al., 2012; Sayasneh et al., 2013a).

The conclusions of this meta-analysis were based on methodologically sound and high-quality evidence having applied Quality Assessment of Diagnostic Accuracy Studies (QUADAS-1) criteria (Kaijser et al., 2014a). In favour of the RMI is that is has been validated more frequently in the literature in comparison to LR2 and SR. However, evidence relating to the diagnostic accuracy of RMI is confounded by the substandard methodological quality and a failure to report potential sources of bias in many of its validation studies (Kaijser et al., 2014b). Most studies that were included in the meta-analysis were small, retrospective, single centre studies that lack information of adequate blinding of the index test (RMI) results to the final outcome after surgery. Such blinding is pivotal in retrospective diagnostic studies that assess relatively subjective index tests such as transvaginal ultrasonography. A lack of blinding in these circumstances can lead to overoptimistic results of diagnostic test performance (Lijmer et al., 1999).

Our meta-analysis focused on a comparison of pooled sensitivity and specificity at the original cut-

off level used for each model. Whilst this is valid from a methodological point of view, to recommend a clinical test only on these criteria is not always correct in practice (Kaijser et al., 2014a). Using different cut-off values for different prediction models (i.e. 10% for LR2, 200 for RMI) does complicate interpretation of the results. In general, for any test to diagnose ovarian cancer, a high sensitivity is the default position, as identifying women with cancer is key to appropriate triage of these patients to specialists in high-volume oncology centres (Kaijser et al., 2013a). This implies one should use a test with a low cut-off. When using a score of 200, RMI misses one in three patients with ovarian cancer. This seems inappropriate as a cut off to triage patients (Timmerman et al., 2014a). Others argue that a triage test should have a high specificity in order to avoid a large number of women with simple cysts being operated on unnecessarily in specialised centres. AUCs are more appropriate statistical method to compare various diagnostic tests, but clinically less meaningful. In all of the IOTA phases the LR2 model had a significantly higher AUC than RMI (Timmerman et al., 2005; Van Holsbeke et al., 2009, 2012; Sayasneh et al., 2013a; Testa et al., 2014). This implies that at every chosen cut-off, LR2 is a better test than RMI. We can illustrate this by supposing both LR2 and RMI have a sensitivity of 95%, in these circumstances the number of false-positive diagnoses of ovarian cancer (benign tumours) will be 23% lower in LR2. Conversely, if both tests would have the same level of specificity (for instance 80%), IOTA LR2 will correctly detect an additional 10% of malignant adnexal tumours (Table III).

Table III. — Diagnostic performance of the logistic regression model 2 (LR2) derived by the International Ovarian Tumour Analysis (IOTA) study and the Risk of Malignancy Index (RMI) using the IOTA 3 dataset (n = 2403).

	RMI	LR2
AUC	0,894	0,925
Specificity if		
Sensitivity 95%	43%	66%
Sensitivity 90%	61%	81%
Sensitivity 85%	71%	86%
Sensitivity if		
Specificity 80%	80%	90%
Specificity 75%	83%	93%
Specificity 70%	86%	94%

LR 2: logistic regression model 2; AUC: area under the receiver-operator characteristics curve (ROC); RMI: risk of malignancy index.

A recent meta-analysis by Nunes et al. (2014) also summarised the currently available evidence concerning the accuracy of IOTA Simple Rules for the diagnosis of ovarian cancer. It includes additional validation studies, but only focused on the accuracy of SR in the masses where they could be applied. They concluded that SR are a reliable triage test in the hands of ultrasound operators with varying levels of expertise with a pooled sensitivity of 93% (95 % CI 90-96%) and specificity of 95% (95 % CI 93-97%) whenever they are applicable (Nunes et al., 2014). In around 20% of women, however, there is a need to undergo additional testing. A two-step strategy using the subjective assessment of ultrasound and colour Doppler findings by a level III expert as a secondary test to characterise those masses where the rules do not apply seems to be the preferred strategy (Table I and II). In our meta-analysis we deliberately decided to summarise results when SR were used as a single test for all tumours by classifying inconclusive cases as malignant in order to obtain a fair and valid comparison to the other prediction models included in the analysis that were designed to classify all tumours (Kaijser et al., 2014a).

Despite its combination of simplicity and excellent performance, two important limitations of the SR are that they cannot be applied in a proportion of masses, and they do not give risk estimates. The IOTA study group is currently working on a different approach to estimate the preoperative risk of cancer in adnexal tumours using a novel risk score based on the ultrasound B- and M-features used in SR. This risk score will be developed on IOTA phase 1 and 2 data (n = 2445), and validated on IOTA phase 3 data. The objectives are to provide excellent discrimination in all tumours in contrast to how SR have been used up to now. This approach should also give well-calibrated risks in both oncology and local community hospitals.

1.3 Multiclass risk prediction of adnexal tumours

The currently available diagnostic tests used for the characterisation of adnexal tumours focus on the presence or absence of cancer. However, optimal patient management also depends on recognising the specific histopathological diagnosis in both benign and malignant conditions (Van Calster et al., 2014). Predicting the specific histopathology of an adnexal mass can lead to avoidance of unnecessary surgery on physiological haemorrhagic cysts, or referral to an appropriate specialist surgeon for an

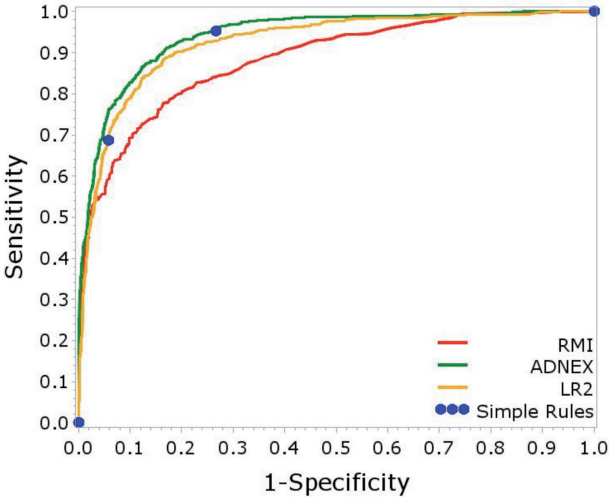


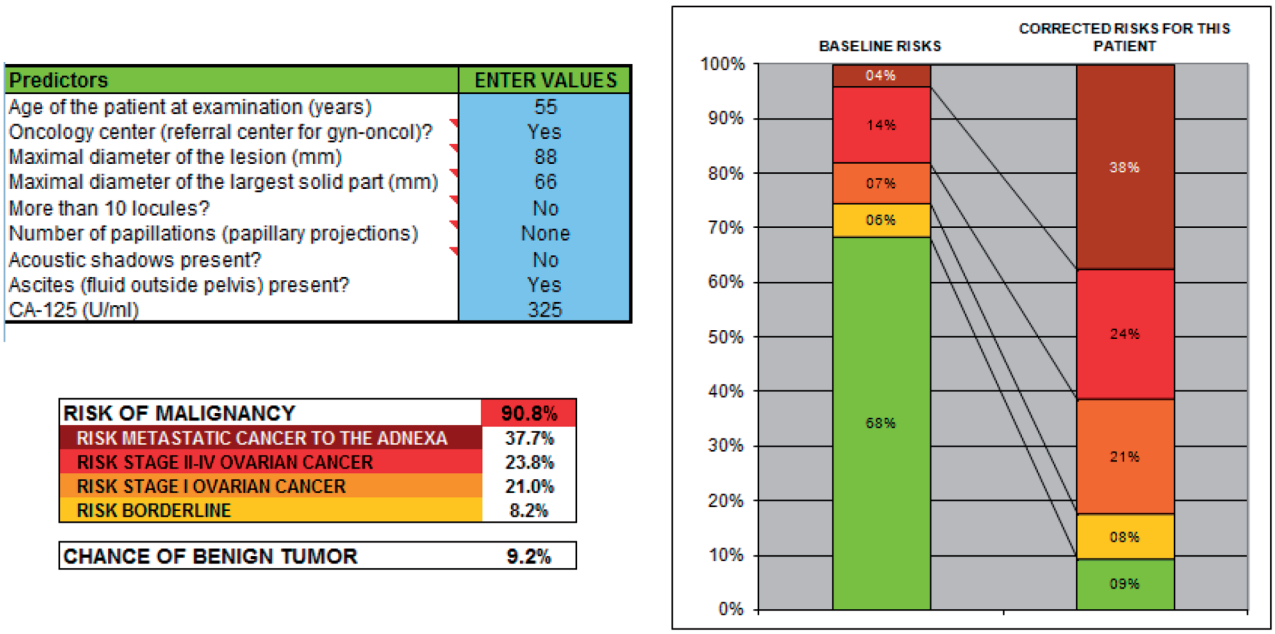
Fig. 3. — Receiver operating characteristic (ROC) curves for the logistic regression model 2 (LR2), Assessment of Different NEoplasias in the AdneXa (ADNEX) model and Risk of Malignancy Index (RMI) with ROC points for the Simple Rules superimposed. The results were obtained using pooled data (n = 2403) from IOTA phase 3.
The left-sided ROC point for Simple Rules represents a situation where the "inconclusive tumours" are classified as benign and the right-sided ROC point a situation where the "inconclusive tumours" are classified as malignant.

endometrioma (Sayasneh et al., 2014). For malignant disease, knowledge of the specific pathology of a lesion may also be critical. For example the need to thoroughly inspect the intestines and perform an appendectomy if a mucinous borderline ovarian tumour (BOT) is predicted (Sayasneh et al., 2014). In IOTA phase 4 we have already shown that clinicians with variable training backgrounds and levels of expertise find it difficult to provide a specific histological diagnosis when using their subjective assessment, especially when discriminating between certain subtypes of malignant disease (BOT, early stage invasive ovarian cancer, and metastatic disease). In this thesis we describe the development and validation of a multiclass risk prediction model ADNEX (Assessment of Different NEoplasias in the AdneXa) using data from IOTA phase 1 to 3 (Van Calster et al., 2014). This polytomous approach to adnexal tumour diagnosis is novel, since established tests like IOTA LR2 or SR only discriminate between cancer and benign conditions. The ADNEX model offered fair to excellent discrimination (AUCs ranging from 0.71 to 0.95) between four different types of ovarian malignancy (BOT, early stage I invasive ovarian cancer, stage II-IV invasive ovarian cancer, and secondary tumours metastasising to the ovaries (i.e. breast, gastro-intestinal tumours, etc)) using risk estimates (Van Calster et al., 2014). Of particular importance is the ability of ADNEX to identify stage I cancer, which the model can discriminate very well from benign tumours and advanced stage cancer and fairly well from BOT and secondary metastatic cancers. On the other hand ADNEX has similar accuracy to IOTA LR2 and SR when validated on IOTA phase 3 data for simple dichotomous risk prediction (benign versus malignant) (Fig. 3) (Testa et al., 2014; Van Calster et al., 2014). ADNEX (AUC 0.94) also performed significantly better than RMI (AUC 0.88) (Timmerman et al., 2014a; Van Calster et al., 2014).

Besides the advantage of polytomous risk prediction, the ADNEX model also provides better calibrated risks for simple dichotomous risk prediction than the IOTA LR2 model when tested on IOTA3 data (Testa et al., 2014). This is likely to be explained by the fact that centre type (oncology centre vs other hospital) is included as a variable in the ADNEX model, and the risk of a malignant tumour is likely to be higher in oncology centres than in other centres, even after adjustments for the characteristics of patients and tumours (Timmerman et al., 2014b).

The use of ADNEX certainly has the potential to further improve and fine-tune management decisions and so reduce the morbidity and mortality associated with adnexal pathology (Van Calster et al., 2014). A sensible option is to use the ADNEX model to first distinguish between a benign and a malignant tumour (for instance based upon a 10% cut-off level) (Van Calster et al., 2015). When the risk exceeds 10% and a malignant tumour is suspected, ADNEX provides four absolute risk estimates for each type of ovarian malignancy based upon the tumour and patient characteristics. Additionally, it also provides the user with baseline risks (based upon the distribution of the four types of malignancies in the combined database of IOTA 1-3). How these predicted patient-risks should be used for triage must be decided on an individual basis (Van Calster et al., 2014). One could interpret absolute risks to determine the most likely malignant pathology, however, due to the low prevalence of certain types of malignancies (i.e. metastatic disease, stage I invasive ovarian cancer) estimated absolute risks for these categories are in general likely to be underestimated, in contrast to those for stage II-IV invasive ovarian cancer (Van Calster et al., 2015). A more practical way to use ADNEX is to also look at the change in the patient-specific risks versus the baseline risks (Van Calster et al., 2015).

In the illustrative example from the ADNEX model shown below (Fig. 4) the relative change in risk is highest for metastatic cancer (i.e. 38%/4%: RR > 9.5). If the relative risk for metastatic cancer exceeds 4 we have shown that the positive predictive value (PPV) is around 32% (Van Calster et al.,



a b

Fig. 4. — Screenshots of the Assessment of Different NEoplasias in the AdneXa (ADNEX) model: probabilities for the five tumour groups displayed in a bar chart (with comparison to baseline probabilities).

2015). The ADNEX model has not been designed to exactly predict the malignant histopathological outcome, but merely serves as a tool to guide clinicians management, for example indicating when other imaging studies (i.e. mammography, colonoscopy or gastroscopy) might be of use (as in the example above), or when conservative surgery (for BOT or stage I invasive ovarian cancer) in young patients might seem reasonable. We feel that the development of a test that characterises the type of malignancy is an important step forward towards individualised patient care (Timmerman et al., 2014a).

The ADNEXmodel contains three clinical (age, serum CA-125 level, type of centre (oncology centre v other hospital), and six ultrasound predictors: maximum diameter of lesion, proportion of solid tissue, more than 10 cyst locules, number of papillary projections, acoustic shadows, and ascites (Fig. 5) (Van Calster et al., 2014). These ultrasound variables are very similar and perhaps simpler to recognise and obtain than variables used in both the IOTA LR2 model and SR (Timmerman et al., 2014a). In contrast to these strategies, ADNEX includes more objective ultrasound features, and does not necessitate the use of colour or power Doppler ultrasonography. Interestingly, the choice of variables used in ADNEX are similar to the RMI. The RMI requires the operator to identify if a mass is multilocular or has solid areas whilst ADNEX asks the examiner to record the number of locules and the size of the solid area if present (Timmerman et al., 2014a). Both the IOTA LR2 model and SR have been shown to be reliable tests in non-expert hands (Nunes et al., 2012, 2013, 2014; Sayasneh et al., 2013a), accordingly ADNEX should also be straightforward to use by any competent ultrasound examiner familiar with IOTA terms and definitions (Timmerman et al., 2014a). An inconvenience that ADNEX shares with existing models to predict ovarian malignancy, such as RMI and the ROMA algorithm, is that optimal predictions can only be made once the CA-125 level is available. ADNEX however, does allow risk calculations to be made without CA-125, however if used in this way predictions for later stage disease will be compromised (Van Calster et al., 2014).

1.4 Serum HE4 and the Risk of Ovarian Malignancy Algorithm for ovarian cancer diagnosis

In most countries biomarkers like serum CA-125 are frequently measured and recommended in protocols to support clinical judgement in patients with symptoms suggestive of ovarian cancer or in those with a confirmed adnexal tumour (Kaijser et al., 2013b).[34] In contrast to diagnostic imaging, they offer the advantage of being objective, relatively cheap and easier to understand. However, these attractions cannot hide the fact that the current test performance of serum CA-125 limits its ability to add clinically to the diagnosis of ovarian cancer as shown in the earlier phases of IOTA (Kaijser et al., 2013b). This led to a search for new complementary biomarkers. At present two new commercial biomarker-based algorithms have been developed;

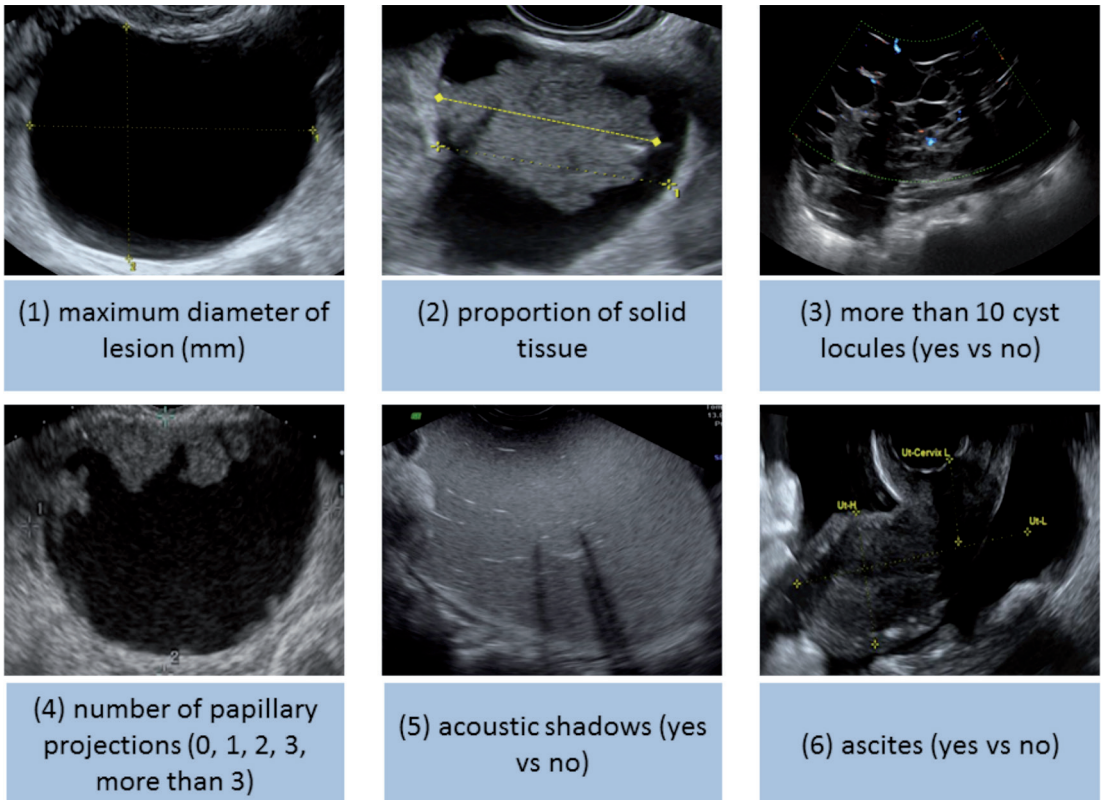


Fig. 5. — Ultrasound variables used in the Assessment of Different NEoplasias in the AdneXa (ADNEX) model derived by the International Ovarian Tumour Analysis (IOTA) study.

the ROMA algorithm (Fujerebio Diagnostics, Malvern, PA) and the Multivariate Index Assay (Vermillion, Austin, TX) (Moore et al., 2009; Ueland et al., 2011; Bristow et al., 2013). Both algorithms have received US Food and Drug Administration (FDA) approval for use in clinical practice.

The Multivariate Index Assay or OVA-1 algorithm incorporates serum levels of multiple biomarkers, including CA-125, transferrin, transthyretin, apolipoprotein A1, and beta 2 microglobulin, and menopausal status, to generate low or high probability of ovarian cancer (Zhang et al., 2004). Two large multicentre trials validated OVA-1 in more than 1000 patients in the USA (Ueland et al., 2011; Bristow et al., 2013). Both studies were included in our meta-analysis (Kaijser et al., 2014a). In these studies OVA-1 achieved comparable diagnostic accuracy with similar sensitivities of 92 and 93% and corresponding specificities of 54 and 43%, respectively (Ueland et al., 2011; Bristow et al., 2013). When replacing serum CA-125 for OVA-1 in the American Congress of Obstetricians and Gynecologists referral guidelines, the sensitivity for cancer and negative predictive value of the College referral guidelines improved whilst drastically decreasing the specificity and positive predictive value (Timmerman et al., 2011; Ware Miller et al., 2011).

The ROMA algorithm published by Moore et al utilises levels of CA-125 and a new emerging epithelial biomarker human-epididymis-protein-4 (HE4) combined with the patient's menopausal status to classify patients as at high or low risk for malignancy (Moore et al., 2009). HE4 was first discovered by Kirchhoff et al. in 1991 and belongs to the "fourdisulfide core" family of proteins, which typically function as proteinase inhibitors (Kirchhoff et al., 1991; Bouchard et al., 2006). Its role as a potential biomarker for ovarian cancer emerged after cDNA comparative hybridization experiments found increased primary expression of HE4 in some ovarian cancers, relative to normal tissues (Schummer et al., 1999). HE4 promotes migration and adhesion of ovarian cancer cells (Lu et al., 2012). It is expressed by a number of normal (breast, colon, lung, male and female genital tract) and malignant tissues (Simmons et al., 2013; Galgano et al., 2006). Amongst malignant tumours, the highest expression levels were observed in ovarian cancer. HE4 expression was positive in 93% of serous, 100% of endometrioid tumours, 50% of clear-cell tumours, but in no mucinous tumours (Drapkin et al., 2005). Since HE4 is overexpressed in ovarian cancers relative to normal tissues, Hellstrom et al. (2003) examined the potential of HE4 as a secreted biomarker for ovarian cancer. As with measurements of serum CA-125, HE4 turned out to be elevated in serum in more than 80% of all invasive epithelial ovarian cancers (Hellstrom et al., 2003; Moore et al.,

2007). When compared to CA-125, it has a similar sensitivity for cancer, but is less frequently elevated in benign conditions such as endometriosis, explaining its higher specificity. This has been confirmed in numerous meta-analyses (Li et al., 2012; Ferraro et al., 2013; Wang et al., 2014; Zhen et al., 2014).

As both markers seemed complementary to each other, this led to the development of the ROMA algorithm (Moore et al., 2009). This commercial test rapidly gained widespread attention, as evidenced by its numerous validation studies throughout the world. Initial multicentre studies conducted in the United States reported an excellent sensitivity of 94% at a pre-set specificity level of 75% (Moore et al., 2009, 2011). Some reports have confirmed this test performance (Bandiera et al., 2011; Kim et al., 2011; Lenhard et al., 2011; Molina et al., 2011; Ruggeri et al., 2011), whilst others have found that ROMA does not outperform established diagnostic tests that incorporate CA-125 or HE4 as single markers (Jacob et al., 2011; Montagnana et al., 2011; Van Gorp et al., 2011). In our systematic review (Kaijser et al., 2014a) we included a total of 18 validation studies (Moore et al., 2009, 2011; Bandiera et al., 2011; Jacob et al., 2011; Kim et al., 2011; Lenhard et al., 2011; Molina et al., 2011; Ruggeri et al., 2011; Montagnana et al., 2011; Partheen et al., 2011; Van Gorp et al., 2011; Anton et al., 2012; Kadija et al., 2012; Karlsen et al., 2012; Presl et al., 2012; Chan et al., 2013; Pitta et al., 2013; Sandri et al., 2013) (n = 5116), but restricted our analysis to only eight of them (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; Pita et al., 2013), because others had systematically excluded BOTs, metastatic and non-epithelial ovarian cancers from their final analysis. Such exclusions lead to biased populations and overoptimistic results of diagnostic test performance. In these eight studies the sensitivity for cancer of ROMA ranged from 57 to 87% and specificity for benign disease from 74 to 95% (Kaijser et al., 2014a).

Within the framework of IOTA we have also compared the ROMA test to different ultrasound-based diagnostic approaches on the same patient populations (Van Gorp et al., 2012; Kaijser et al., 2013c). In the first study including a total of 374 patients the subjective assessment of level III expert ultrasound examiners remained superior in discriminating malignant from benign ovarian masses when compared to both ROMA and RMI. The pre and postmenopausal populations generated similar results. Interestingly, in this study the RMI also had a significantly higher AUC than ROMA. In the second study, as presented in this thesis, we describe our findings of a single-centre diagnostic accuracy study (n = 360) comparing the ROMA algorithm to the use of IOTA LR2 in the hands of examiners with different levels of expertise on a population of 360 patients with known ovarian tumours. IOTA LR2 was a significantly better and more reliable test than ROMA (Kaijser et al., 2013c). In premenopausal women ROMA missed 1/3 of all ovarian malignancies, whilst LR2 only missed 1 in 20 ovarian cancers. Additionally, LR2 also proved to be better able to diagnose specific subtypes of ovarian carcinoma. The retrospective evaluation of test performance of both strategies in this study is not very likely to have introduced bias, as patient outcomes were blinded when calculating predicted patient-risks of cancer. The findings of both studies suggest that whenever ultrasonography is available to clinicians with different levels of ultrasound expertise, and IOTA models are used in the correct and appropriate way both serum CA-125 and HE4 do not contribute significantly to the diagnosis of ovarian cancer.

We also evaluated the potential role of serum HE4 and the ROMA test as a secondary test in the group of adnexal tumours (about 8% of the total) that can not be classified with a high level of confidence by experienced Level III sonographers. Amongst unclassifiable cases, serous and mucinous cystadenomas/cystadenofibromas, fibromas, rare benign tumours and borderline tumours are often overrepresented (Valentin et al., 2011). However, both serum HE4 and ROMA had very low discriminatory capacity in this very specific group of adnexal tumours with poor AUCs of 0.536 and 0.565, respectively (Kaijser et al., 2014c).

1.5 Imaging techniques to diagnose adnexal tumours

TVS is accepted as the most appropriate initial imaging investigation to identify and characterise any mass found in women suspected of having adnexal pathology (Dodge et al., 2012). Compared to other imaging techniques, it has the great advantage that it is dynamic, interactive, allows site-specific tenderness to be assessed, along with mobility of the tumour with respect to adjacent structures in the pelvis (Testa et al., 2009). Other imaging techniques such as computed-tomography (CT), magnetic resonance imaging (MRI), and [18F]-fluorodeoxyglucose positron emission tomography ([18F]-FDG-PET) are also being used in clinical practice to classify the nature of adnexal masses prior to surgery (Kaijser et al., 2014d). CT and FDG-PET however both lack accuracy for

preoperative diagnosis and are more useful tools for staging of known malignant disease and the assessment of ovarian cancer recurrence. MRI in turn may have a limited role to play in masses characterised as being "difficult to classify" with ultrasonography (Kinkel et al., 2005; Anthoulakis et al., 2013). However, in these studies there was no agreement on what qualified as a "difficult mass". Furthermore they did not report the quality of TVS and experience level of the examiners (Kaijser et al., 2014d). Unique to MRI compared with conventional grey scale and colour Doppler ultrasound is its ability to characterise adnexal lesions by measuring their cellularity, a technique called diffusion-weighted imaging (DWI) (Le Bihan, 1995) and by allowing a (semi-) quantitative reflection of lesion microvascularisation using dynamic contrast-enhanced (DCE) MRI (Kaijser et al., 2014d; Thomassin-Naggara et al., 2012; Bernardin et al., 2012). The addition of these sequences to conventional MRI has led to increased diagnostic accuracy. Both sequences provided new criteria to those used in conventional MRI protocols to describe "difficult" tumours using TVS and has led to the development of a novel $A_{DNEX}MR_{SCORING}$system (Thomassin-Naggara et al., 2013). This classification system aimed to improve the standardisation of MRI reports and has showed encouraging discrimination of these so-called "difficult" adnexal tumours. However, independent external validation is first required to confirm its initial promising test accuracy results. The IOTA risk models or Simple Rules might prove useful in this sense. In the UK at present in many protocols MRI is advised to help classify masses with an intermediate Risk of Malignancy (RMI)-score of 25-250. However, if IOTALR2-based triage is used instead, the total number of women with "difficult to classify" tumours sent for MRI would be decreased substantially (17.8% versus 31.4%) (Van Calster et al., 2012).

Conclusion

After more than 15 years of extensive diagnostic research in which the main IOTA approaches have been compared to the current reference standard test (RMI) in a large number of patients, in various clinical environments, in different countries, and by examiners with different backgrounds and clinical expertise, we feel it has become evident that an "evidence-based" approach to the characterisation of adnexal masses needing surgical intervention should include IOTA LR2 or Simple Rules when level III ultrasound expertise is not available (Fig. 6). In the near future we think ADNEX will

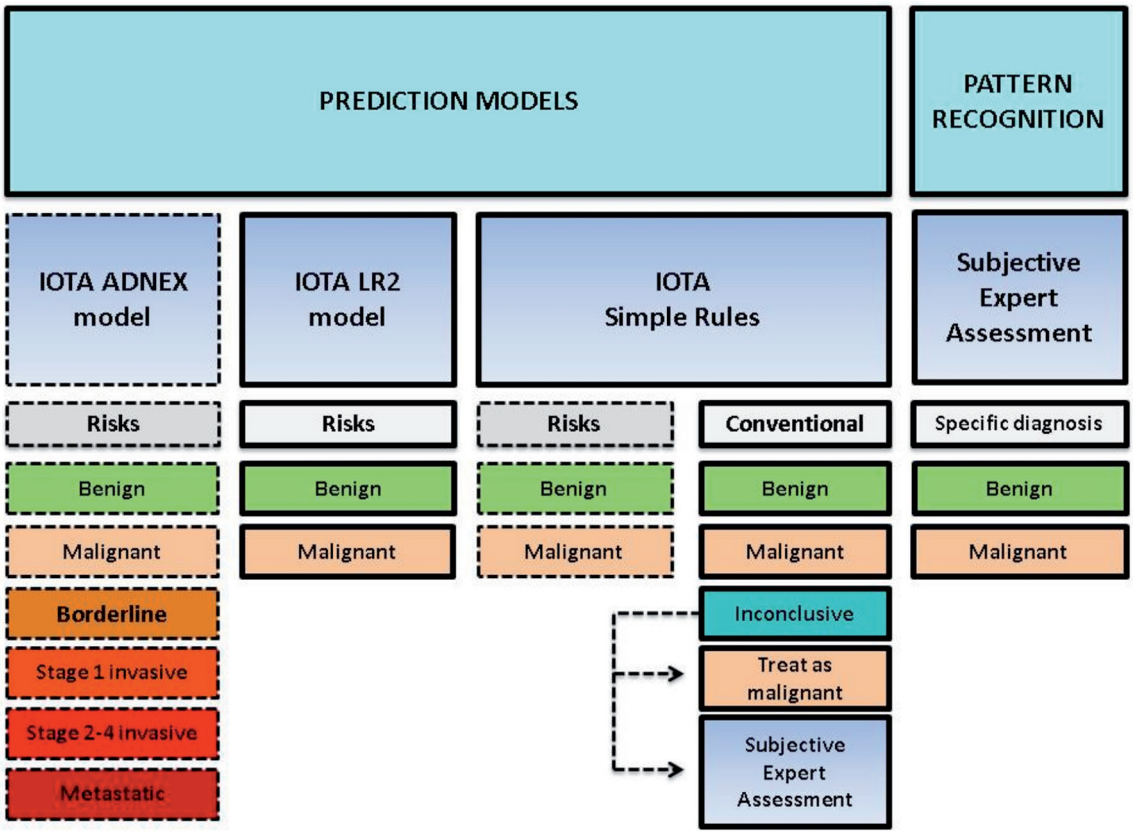


Fig. 6. – Flowchart showing the optimal "evidence-based" approaches for the assessment of women with adnexal masses that require surgery to estimate preoperative risk of malignancy after critical appraisal of available evidence.

supersede LR2 and when a prediction model is used this will become the approach of choice for classifying adnexal tumours.

To facilitate use of IOTA models in clinical practice and to support clinical decision making, the IOTA group has developed simple, user-friendly mobile applications (apps) for IOTA LR1, LR2, and SR for both IOS ("iotamodels") and Android smartphones or tablets. Very recently, it has also launched the ADNEX app for IOS and Android ("iota2014") and a free to download web-based interface available on http://www.iotagroup.org/adnexmodel.

Clinical Implications

Their use should improve the existing management, guidance and triage of patients with adnexal tumours for subspecialist care. If theoretically, the IOTA models were implemented instead of the RMI in Belgium and The Netherlands alone, an additional 160 and 250 patients, respectively, could have been correctly diagnosed with ovarian cancer on a yearly basis (www.cijferstegenkanker.nl and – www.kankerregister.org). Such an IOTA-based triage of patients would thus prevent many unnecessary secondary surgical staging procedures in false positive cases, and will have a positive influence on the long-term prognosis in patients with cancer.

It is therefore to be hoped that evidence-based guidelines by professional societies on management of women with suspected ovarian cancer will soon be updated to reflect these improvements in preoperative diagnosis as described in detail in this thesis. Based upon all the available evidence there is currently no need for women at high risk of ovarian cancer not to have surgery performed by the right surgeon, in the right place, at the right time.

Topics of Future Research for the IOTA study

1. External validation of the multiclass IOTA ADNEX model in the hands of operators with different training and experience
2. Second-stage tests in tumours that are "difficult to classify" using ultrasonography (value of new biomarkers, pelvic MRI with perfusion and diffusion-weighted imaging ($A_{DNEX}MR_{SCORING}$ system) and 3-D power Doppler imaging.
3. Improve the calibration of established prediction models for ovarian cancer.
4. Investigate the use of Clinical utility (Net Benefit): an alternative, clinically more interpretable measure to compare usefulness of different prediction models for ovarian cancer.

The overall aim of the IOTA study was to develop the optimal algorithm or model for the classification and management (minimal invasive surgery, subspecialist treatment, conservative management, additional testing) of all types of adnexal pathology. All the IOTA studies that have been discussed in this thesis have been directed towards the use of diagnostic tests and models to predict malignancy in ovarian tumours that have subsequently been surgically removed in order to provide a clear histological endpoint. There are still several themes for future diagnostic research on operated masses that can be discerned:

In addition, there are various other, and maybe more important challenges to overcome for the IOTA consortium in order to be able to definitively advise patients on the relative merits of both surgery and expectant management.

Need for true interventional trials

A limitation of most diagnostic research is that prediction models are not directly applied when scanning patients with adnexal tumours. Instead, in each study, various clinical and ultrasound variables are prospectively collected from each patient and later incorporated into these models to calculate and compare test performance. However, what we really need are prospective and preferably randomised interventional studies that directly apply diagnostic tests to patients with adnexal pathology. Only this type of research will truly establish whether promising new diagnostic tools like the ultrasound-based IOTA approaches will have a positive influence on both clinical management and patient outcomes. Two randomised interventional studies in the UK have compared IOTA Simple Rules and LR2 based protocols to the use of RMI in patients with adnexal tumours (IOTA phase 4b and ISRCTN89034131: registered at http://www.controlled-trials.com/ISRCTN89034131. Both trials have ended their recruitment in 2014 and final results are to be awaited.

Studying the long-term outcome after conservative management of cysts classified as benign: IOTA phase 5

The decision to operate on an ovarian mass depends on a number of factors. These include the subjective characterisation of the mass using ultrasound, the use of simple prediction models, the age of the patient, co-morbidity, family history, the serum CA-125 level and the presence of symptoms such as pain. The management of cysts that are not removed

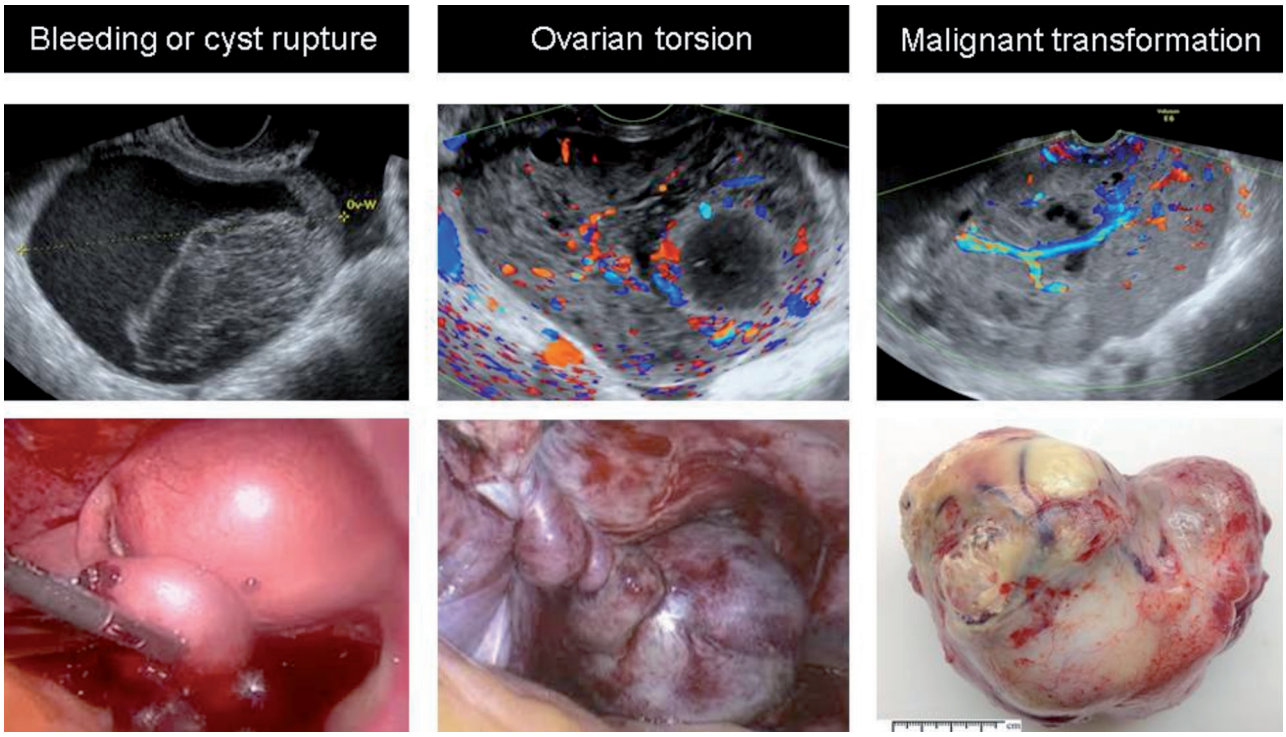


Fig. 7. — Possible complications after conservative management of benign classified adnexal cysts

surgically is not evidence-based and often subject to wide variation. In the absence of rigorous follow up data, we do not know how many false negative results with respect to cancer are associated with these cysts, or if they sometimes undergo malignant transformation, bleeding, rupture or torsion (Fig. 7).

Because the natural history of such ovarian masses is not known, and because of the fear of "missing" ovarian cancer, many ovarian masses are currently surgically removed in asymptomatic women, even if they do not show signs of malignancy. Developing new insights into the natural history of benign looking conservatively managed ovarian masses would potentially change the management of thousands of women, by avoiding surgery or even further surveillance for some and detecting cancer earlier or even preventing it for others. This type of knowledge will only be gained by long-term systematic follow up of a large cohort of ovarian cysts. To date, no research has rigorously investigated the long-term behaviour of such masses. The long-term follow-up arm of IOTA phase 5 which started in 2012 is the first international study of its kind. Its results are likely to lead to significantly reduced intervention and associated morbidity and mortality in some, and better directed intervention in women with cancer and those with lesions that have a high risk of becoming malignant in the short to medium term.

Development and validation of a new evidence-based, multilevel, multimodal risk-scoring tool for women with suspected ovarian cancer

A limitation of current diagnostic research for ovarian cancer is its strong focus on specific data-modalities and specific patient care pathways (screening, primary care, secondary/tertiary care). There is an abundance of literature on the use of ultrasound-based tests to classify ovarian masses in secondary/tertiary care, and on the use of symptoms or symptom-based indices to triage symptomatic women at high risk of cancer in a primary care setting. But there is only limited "cross-fertilisation". A single multimodal approach that can be integrated into different healthcare settings or future proof new care pathways (screening for ovarian cancer) is challenging but would certainly be of benefit to the patient, as it will streamline and optimise current diagnostic pathways.

This relates to the ultimate goal of the ROCkeTS (Refining Ovarian Cancer Test Accuracy Scores) project starting in October 2014 (http://www.nets.nihr.ac.uk/projects/hta/131301). By using an amalgated dataset from the International Ovarian Tumour Analysis (IOTA) study, the United Kingdom Ovarian Cancer Population Study (UKOPS) and the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) the ROCkeTS projects aims:

1) To derive and validate new and refined multimodal (symptoms, CA-125, new biomarkers,

ultrasound) risk tools (build on existing risk scores) that estimate the probability of having ovarian cancer for women with symptoms suggestive of cancer that are applicable to both primary and secondary care.

2) To investigate potential new second-stage tests in asymptomatic screen-positive women. If multimodal screening (longitudinal change of CA-125 as a primary test combined with transvaginal ultrasonography as a secondary test) in UKCTOCS will save lives the IOTA Simple Rules or ADNEX model might be important candidates to further improve screening strategies for ovarian cancer.

Acknowledgements

Dirk Timmerman is Senior Clinical Investigator of the Research Foundation - Flanders (Belgium) (FWO). Tom Bourne is 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 author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. The IOTA studies are supported by the Flemish Government: Research Foundation – Flanders (FWO) project G049312N, and Flanders' Agency for Innovation by Science and Technology (IWT) project IWT-TBM 070706-IOTA3.

References

Alcázar JL, Pascual MÁ, Olartecoechea B et al. IOTA simple rules for discriminating between benign and malignant adnexal masses: prospective external validation. Ultrasound ObstetGynecol. 2013;42:467-71.

Ameye L, Timmerman D, Valentin L et al. Clinically oriented three-step strategy for assessment of adnexal pathology. Ultrasound Obstet Gynecol. 2012;40:582-91.

Anthoulakis C, Nikoloudis N. Pelvic MRI as the "gold standard" in the subsequent evaluation of ultrasound-indeterminate adnexal lesions: a systematic review. Gynecol Oncol. 2014;132:661-8.

Anton C, Carvalho FM, Oliveira EI et al. A comparison of CA125, HE4, risk ovarian malignancy algorithm (ROMA), and risk malignancy index (RMI) for the classification of ovarian masses. Clinics (Sao Paulo) 2012;67:437-41.

Bandiera E, Romani C, Specchia C et al. Serum human epididymis protein 4 and risk for ovarian malignancy algorithm as new diagnostic and prognostic tools for epithelial ovarian cancer management. Cancer Epidemiol Biomarkers Prev. 2011;20:2496–2506.

Bernardin L, Dilks P, Liyanage S et al. Effectiveness of semi-quantitative multiphase dynamic contrast-enhanced MRI as a predictor of malignancy in complex adnexal masses: radiological and pathological correlation. Eur Radiol. 2012;22:880-90.

Bouchard D, Morisset D, Bourbonnais Y et al. Proteins with whey-acidic-protein motifs and cancer. Lancet Oncol. 2006;7:167-74.

Bristow RE, Smith A, Zhang Z et al. Ovarian malignancy risk stratification of the adnexal mass using a multivariate index assay. Gynecol Oncol. 2013;128:252-9.

Campbell S. Ovarian cancer: role of ultrasound in preoperative diagnosis and population screening. Ultrasound Obstet Gynecol. 2012;40:245-54.

Chan KK, Chen CA, NamJH et al. The use of HE4 in the prediction of ovarian cancer in Asian women with a pelvic mass. Gynecol Oncol. 2013;128:239-44.

Dodge JE, Covens AL, Lacchetti C et al; The Gynecology Cancer Disease Site Group. Preoperative identification of a suspicious adnexal mass: a systematic review and meta-analysis. Gynecol Oncol. 2012;126:157-66.

Drapkin R, Von Horsten HH, Lin Y et al. Human epididymis protein 4 (HE4) is a secreted glycoprotein that is overexpressed by serous and endometrioid ovarian carcinomas. Cancer Res. 2005;65:2162-9.

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

Fathallah K, Huchon C, Bats AS et al. External validation of simple ultrasound rules of Timmerman on 122 ovarian tumors. Gynecol Obstet Fertil. 2011;39:477-81.

Ferraro S, Braga F, Lanzoni M et al. Serum human epididymis protein 4 vs carbohydrate antigen 125 for ovarian cancer diagnosis: a systematic review. J Clin Pathol. 2013;66:273-81.

Galgano MT, Hampton GM, Frierson HF Jr. Comprehensive analysis of HE4 expression in normal and malignant human tissues. Mod Pathol. 2006;19:847-53.

Hartman CA, Juliato CR, Sarian LO et al. Ultrasound criteria and CA 125 as predictive variables of ovarian cancer in women with adnexal tumors. Ultrasound Obstet Gynecol. 2012; 40:360-6.

Hellstrom I, Raycraft J, Hayden-Ledbetter M et al. The HE4 (WFCD2) protein is a biomarker for ovarian carcinoma. Cancer Res. 2003;63:3695-700.

Jacob F, Meier M, Caduff R et al. No benefit from combining HE4 and CA125 as ovarian tumor markers in a clinical setting. Gynecol Oncol. 2011;121:487-91.

Jacobs I, Oram D, Fairbanks J et al. A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer. Br J Obstet Gynaecol.1990;97:922-9.

Kadija S, Stefanovic A, Jeremic K et al. The utility of human epididymal protein 4, cancer antigen 125, and risk for malignancy algorithm in ovarian cancer and endometriosis. Int J Gynecol Cancer. 2012;22:238-44.

Kaijser J, Bourne T, De Rijdt S et al. Key findings from the International Ovarian Tumor Analysis (IOTA) study: an approach to the optimal ultrasound based characterization of adnexal pathology. AJUM. 2012;15:82-6.

Kaijser J, Bourne T, Valentin L et al. Improving strategies for diagnosing ovarian cancer: a summary of the International Ovarian Tumor Analysis (IOTA) studies. Ultrasound ObstetGynecol. 2013a;41:9-20.

Kaijser J, Van Gorp T, Sayasneh A et al. Differentiating stage I epithelial ovarian cancer from benign disease in women with adnexal tumors using biomarkers or the ROMA algorithm. Gynecol Oncol. 2013b;130:398-9.

Kaijser J, Van Gorp T, Van Hoorde K et al. 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. 2013c;129:377-83.

Kaijser J, Sayasneh A, Van Hoorde K et al. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. Hum Reprod Update. 2014a;20:449-62.

Kaijser J, Sayasneh A, Van Hoorde K et al. Preoperatief differentiëren tussen benigne en maligne adnex pathologie: nieuwe evidence-based adviezen voor de dagelijkse praktijk. NTOG. 2014b;117:274-82.

Kaijser J, Van Gorp T, Smet ME et al. Are serum HE4 or ROMA scores useful to experienced examiners to improve charac-

terization of adnexal masses after transvaginal ultrasonography? Ultrasound Obstet Gynecol. 2014c;43:89-97.
Kaijser J, Vandecaveye V, Deroose CM et al. Imaging techniques for the presurgical diagnosis of adnexal tumours. Best Pract Res Clin Obstet Gynecol. 2014d;28:683-95.
Karlsen MA, Sandhu N, Høgdall C et al. Evaluation of HE4, CA125, risk of ovarian malignancy algorithm (ROMA) and risk of malignancy index (RMI) as diagnostic tools of epithelial ovarian cancer in patients with a pelvic mass. Gynecol Oncol. 2012;127:379-83.
Kim YM, Whang DH, Park J et al. Evaluation of the accuracy of serum human epididymis protein 4 in combination with CA125 for detecting ovarian cancer: a prospective case-control study in a Korean population. Clin Chem Lab Med. 2011;49:527-34.
Kinkel K, Lu Y, Mehdizade A et al. Indeterminate ovarian mass at US: incremental value of second imaging test for characterization--meta-analysis and Bayesian analysis. Radiology. 2005;236:85-94.
Kirchhoff C, Habben I, Ivell R et al. A major human epididymis-specific cDNA encodes a protein with sequence homology to extracellular proteinase inhibitors. Biol Reprod. 1991;45:350-7.
Le Bihan D. Molecular diffusion, tissue microdynamics and microstructure. NMR Biomed. 1995;8:375-86.
Lenhard M, Stieber P, Hertlein L et al. The diagnostic accuracy of two human epididymis protein 4 (HE4) testing systems in combination with CA125 in the differential diagnosis of ovarian masses. Clin Chem Lab Med. 2011;49:2081-8.
Li F, Tie R, Chang K et al. Does risk for ovarian malignancy algorithm excel human epididymis protein 4 and CA125 in predicting epithelial ovarian cancer: a meta-analysis. BMC Cancer. 2012;12:258.
Lijmer JG, Mol BW, Heisterkamp S et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;11:1061-6.
Lu R, Sun X, Xiao R et al. Human epididymis protein 4 (HE4) plays a key role in ovarian cancer cell adhesion and motility. Biochem Biophys Res Commun. 2012;419:274-80.
Molina R, Escudero JM, Augé JM et al. HE4 a novel tumour marker for ovarian cancer: comparison with CA 125 and ROMA algorithm in patients with gynaecological diseases. Tumor Biol. 2011;32:1087-95.
Montagnana M, Danese E, Ruzzenente O et al. The ROMA (Risk of Ovarian Malignancy Algorithm) for estimating the risk of epithelial ovarian cancer in women presenting with pelvic mass: is it really useful? Clin Chem Lab Med. 2011;49:521-5.
Moore RG, Brown AK, Miller MC et al. The use of multiple novel tumor biomarkers for the detection of ovarian carcinoma in patients with a pelvic mass. Gynecol Oncol. 2007;108:402-8.
Moore RG, McMeekin DS, Brown AK et al. A novel multiple marker bioassay utilizing HE4 and CA125 for the prediction of ovarian cancer in patients with a pelvic mass. GynecolOncol 2009;112:40-6.
Moore RG, Miller MC, Disilvestro P et al. Evaluation of the diagnostic accuracy of the risk of ovarian malignancy algorithm in women with a pelvic mass. Obstet Gynecol. 2011;118(2 part 1):280-8.
Nunes N, Yazbek J, Ambler G et al. Prospective evaluation of the IOTA logistic regression model LR2 for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol. 2012;40:355-9.
Nunes N, Ambler G, Foo X et al. Use of IOTA simple rules for diagnosis of ovarian cancer: meta-analysis. Ultrasound Obstet Gynecol. 2014;44:503-14.
Nunes N, Ambler G, Hoo WL et al. 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. 2013;23:1583-9.
Partheen K, Kristjansdottir B, Sundfeldt K. Evaluation of ovarian cancer biomarkers HE4 and CA-125 in women presenting with a suspicious cystic ovarian mass. J Gynecol Oncol. 2011;22:244-52.
Pitta DD, Sarian LO, Barreta A et al. Symptoms, CA125 and HE4 for the preoperative prediction of ovarian malignancy in Brazilian women with ovarian masses. BMC Cancer. 2013;13:423.
Presl J, Kučera R, Topolčan O et al. HE4 a biomarker of ovarian cancer. Ceska Gynekol. 2012;77:445-9.
Ruggeri G, Bandiera E, Zanotti L et al. HE4 and epithelial ovarian cancer: comparison and clinical evaluation of two immunoassays and a combination algorithm. Clin Chim Acta. 2011;412:1447-53.
Sandri MT, Bottari F, Franchi D et al. Comparison of HE4, CA125 and ROMA algorithm in women with a pelvic mass: correlation with pathological outcome. Gynecol Oncol. 2013;128:233-8.
Sayasneh A, Wynants L, Preisler J et al. Multicentre external validation of IOTA prediction models and RMI by operators with varied training. Br J Cancer. 2013a;108:2448-54.
Sayasneh A, Kaijser J, Preisler J et al. A multicenter prospective external validation of the diagnostic performance of IOTA simple descriptors and rules to characterize ovarian masses. Gynecol Oncol. 2013b;130:140-6.
Sayasneh A, Kaijser J, Preisler J et al. The accuracy of ultrasonography performed by examiners with varied training and experience to predict the specific pathology of adnexal masses. Ultrasound ObstetGynecol 2014 Oct 1. doi: 10.1002/uog.14675.
Schummer M, Ng WV, Bumgarner RE et al. Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene. 1999;238:375-85.
SIGN 135. Management of epithelial ovarian cancer. A national clinical guideline, Issued November 2013. Available at: http://www.sign.ac.uk/pdf/sign135.pdf . Accessed October 31, 2014.
Simmons AR, Baggerly K, Bast RC Jr. The emerging role of HE4 in the evaluation of epithelial ovarian and endometrial carcinomas. Oncology. 2013;27:548-56.
Sladkevicius P, Valentin L. Inter-observer agreement in describing the ultrasound appearance of adnexal masses and in calculating the risk of malignancy using logistic regression models. Clin Cancer Res. 2015;21:594-601.
Tantipalakorn C, Wanapirak C, Khunamornpong S et al. IOTA simple rules in differentiating between benign and malignant ovarian tumors. Asian Pac J Cancer Prev. 2014;15:5123-6.
Testa AC, Bourne TH. Characterising pelvic masses using ultrasound. Best Pract Res Clin Obstet Gynaecol. 2009; 23:725-38.
Testa A, Kaijser J, Wynants L et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicentre international IOTA study. Br J Cancer. 2014;111:680-8.
Thomassin-Naggara I, Balvay D, Aubert E et al. Quantitative dynamic contrast-enhanced MR imaging analysis of complex adnexal masses: a preliminary study. Eur Radiol. 2012; 22:738-45.
Thomassin-Naggara I, Aubert E, Rockall A et al. Adnexal masses: development and preliminary validation of an MR imaging scoring system. Radiology. 2013;267:432-43.
Timmerman D, Valentin L, Bourne TH et al. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. Ultrasound Obstet Gynecol. 2000;16:500-5.
Timmerman D, Testa AC, Bourne T et al. International Ovarian Tumor Analysis Group. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. J Clin Oncol. 2005;23:8794-801.
Timmerman D, Van Calster B, Jurkovic D et al. Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors. J Clin Oncol. 2007;25:4194-200.

Timmerman D, Testa AC, Bourne T et al. Simple ultrasound-based rules for the diagnosis of ovarian cancer. Ultrasound Obstet Gynecol. 2008;31:681-90.

Timmerman D, Van Calster B, Testa AC et al. Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. Ultrasound Obstet Gynecol. 2010a; 36:226-34.

Timmerman D, Ameye L, Fischerova D et al. Simple ultrasound rules to distinguish between benign and malignant adnexal masses before surgery: prospective validation by IOTA group. BMJ. 2010b;341:c6839.

Timmerman D, Van Calster B, Vergote I et al. Performance of the American College of Obstetricians and Gynecologists' ovarian tumor referral guidelines with a multivariate index assay. ObstetGynecol. 2011;118:1179-81.

Timmerman D, Van Calster B, Kaijser J et al. Reply: The ADNEX model is based on simple to use variables and performs better than the conventional Risk of Malignancy Index (RMI) (2014a). Accessed at http://www.bmj.com/content/349/bmj.g5920/rr/778859.

Ueland FR, Desimone CP, Seamon LG et al. Effectiveness of a multivariate index assay in the preoperative assessment of ovarian tumors. Obstet Gynecol. 2011;117:1289-97.

Valentin L, Ameye L, Jurkovic D et al. Which extrauterine pelvic masses are difficult to correctly classify as benign or malignant on the basis of ultrasound findings and is there a way of making a correct diagnosis? Ultrasound Obstet Gynecol. 2006;27:438-44.

Valentin L, Jurkovic D, Van Calster B et al. Adding a single CA-125 measurement to ultrasound performed by an experienced examiner does not improve preoperative discrimination between benign and malignant adnexal masses. Ultrasound Obstet Gynecol. 2009;34:345-54.

Valentin L, Ameye L, Savelli L et al. Adnexal masses difficult to classify as benign or malignant using subjective assessment of gray-scale and Doppler ultrasound findings: logistic regression models do not help. Ultrasound Obstet Gynecol. 2011;38:456-65.

Van Calster B, Timmerman D, Bourne T et al. Discrimination between benign and malignant adnexal masses by specialist ultrasound examination versus serum CA-125. J Natl Cancer Inst. 2007;99:1706-14.

Van Calster B, Valentin L, Van Holsbeke C et al. Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. BMC Med Res Methodol. 2010;10:96.

Van Calster B, Timmerman D, Valentin L et al. Triaging women with ovarian masses for surgery: observational diagnostic study to compare RCOG guidelines with an International Ovarian Tumour Analysis (IOTA) group protocol. BJOG. 2012;119:662-71.

Van Calster B, Valentin L, Testa A et al. Diagnosing ovarian cancer using the ADNEX risk model from the International Ovarian Tumour Analysis group: differentiating between benign, borderline, stage I invasive, advanced stage invasive, and secondary metastatic tumours. BMJ. 2014;349: g5920.

Van Calster B, Van Hoorde K, Froyman W, et al. Practical guidance for applying the ADNEX model from the IOTA group to discriminate between different subtypes of adnexal tumors. Facts Views Vis Obgyn 2015;7:32-41.

Van Gorp T, Cadron I, Despierre E et al. HE4 and CA125 as a diagnostic test in ovarian cancer: prospective validation of the Risk of Ovarian Malignancy Algorithm. Br J Cancer. 2011;104:863-70.

Van Gorp T, Veldman J, Van Calster B et al. Subjective assessment by ultrasound is superior to the risk of malignancy index (RMI) or the risk of ovarian malignancy algorithm (ROMA) in discriminating benign from malignant adnexal masses. Eur J Cancer. 2012;48:1649-56.

Van Holsbeke C, Van Calster B, Testa AC et al. Prospective internal validation of mathematical models to predict malignancy in adnexal masses: results from the international ovarian tumor analysis study. Clin Cancer Res. 2009;15:684-91.

Van Holsbeke C, Van Calster B, Bourne T et al. External Validation of diagnostic models to estimate the risk of malignancy in adnexal masses. Clin Cancer Res. 2012; 18:815-25.

Wang J, Gao J, Yao H et al. Diagnostic accuracy of serum HE4, CA125 and ROMA in patients with ovarian cancer: a meta-analysis. Tumour Biol. 2014;35:6127-38.

Ware Miller R, Smith A, DeSimone CP et al. Performance of the American College of Obstetricians and Gynecologists' ovarian tumor referral guidelines with a multivariate index assay. Obstet Gynecol. 2011;117:1298-306.

Zhang Z, Bast RC Jr, Yu Y et al. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res. 2004;64:5882-90.

Zhen S, Bian LH, Chang LL et al. Comparison of serum human epididymis protein 4 and carbohydrate antigen 125 as markers in ovarian cancer: a meta-analysis. Mol Clin Oncol. 2014;2:559-66.