Risk-prediction models in postmenopausal patients with symptoms of suspected ovarian cancer in the UK (ROCkeTS): a multicentre, prospective diagnostic accuracy study

*Sudha Sundar, Ridhi Agarwal, Clare Davenport, Katie Scandrett, Susanne Johnson, Partha Sengupta, Radhika Selvi-Vikram, Fong Lien Kwong, Sue Mallett, Caroline Rick, Sean Kehoe, Dirk Timmerman, Tom Bourne, Ben Van Calster, Hilary Stobart, Richard D Neal, Usha Menon, Alex Gentry-Maharaj, Lauren Sturdy, Ryan Ottridge, Jon Deeks, for the ROCkeTS collaborators**

Summary

Background Multiple risk-prediction models are used in clinical practice to triage patients as being at low risk or high risk of ovarian cancer. In the ROCkeTS study, we aimed to identify the best diagnostic test for ovarian cancer in symptomatic patients, through head-to-head comparisons of risk-prediction models, in a real-world setting. Here, we report the results for the postmenopausal cohort.

Methods In this multicentre, prospective diagnostic accuracy study, we recruited newly presenting female patients aged 16–90 years with non-specific symptoms and raised CA125 or abnormal ultrasound results (or both) who had been referred via rapid access, elective clinics, or emergency presentations from 23 hospitals in the UK. Patients with normal CA125 and simple ovarian cysts of smaller than 5 cm in diameter, active non-ovarian malignancy, or previous ovarian malignancy, or those who were pregnant or declined a transvaginal scan, were ineligible. In this analysis, only postmenopausal participants were included. Participants completed a symptom questionnaire, gave a blood sample, and had transabdominal and transvaginal ultrasounds performed by International Ovarian Tumour Analysis consortium (IOTA)-certified sonographers. Index tests were Risk of Malignancy 1 (RMI1) at a threshold of 200, Risk of Malignancy Algorithm (ROMA) at multiple thresholds, IOTA Assessment of Different Neoplasias in the Adnexa (ADNEX) at thresholds of 3% and 10%, IOTA SRRisk model at thresholds of 3% and 10%, IOTA Simple Rules (malignant *vs* benign, or inconclusive), and CA125 at 35 IU/mL. In a post-hoc analysis, the Ovarian Adnexal and Reporting Data System (ORADS) at 10% was derived from IOTA ultrasound variables using established methods since ORADS was described after completion of recruitment. Index tests were conducted by study staff masked to the results of the reference standard. The comparator was RMI1 at the 250 threshold (the current UK National Health Service standard of care). The reference standard was surgical or biopsy tissue histology or cytology within 3 months, or a self-reported diagnosis of ovarian cancer at 12 month follow-up. The primary outcome was diagnostic accuracy at predicting primary invasive ovarian cancer versus benign or normal histology, assessed by analysing the sensitivity, specificity, C-index, area under receiver operating characteristic curve, positive and negative predictive values, and calibration plots in participants with conclusive reference standard results and available index test data. This study is registered with the International Standard Randomised Controlled Trial Number registry (ISRCTN17160843).

Findings Between July 13, 2015, and Nov 30, 2018, 1242 postmenopausal patients were recruited, of whom 215 (17%) had primary ovarian cancer. 166 participants had missing, inconclusive, or other reference standard results; therefore, data from a maximum of 1076 participants were used to assess the index tests for the primary outcome. Compared with RMI1 at 250 (sensitivity 82·9% [95% CI 76·7 to 88·0], specificity 87·4% [84·9 to 89·6]), IOTA ADNEX at 10% was more sensitive (difference of –13·9% [–20·2 to –7·6], p<0·0001) but less specific (difference of 28·5% [24·7 to 32·3], p<0·0001). ROMA at 29·9 had similar sensitivity (difference of –3·6% [–9·1 to 1·9], p=0·24) but lower specificity (difference of 5·2% [2·5 to 8·0], p=0·0001). RMI1 at 200 had similar sensitivity (difference of –2·1% [–4·7 to 0·5], p=0·13) but lower specificity (difference of 3·0% [1·7 to 4·3], p<0·0001). IOTA SRRisk model at 10% had similar sensitivity (difference of –4·3% [–11·0 to –2·3], p=0·23) but lower specificity (difference of 16·2% [12·6 to 19·8], p<0·0001). IOTA Simple Rules had similar sensitivity (difference of –1·6% [–9·3 to 6·2], p=0·82) and specificity (difference of –2·2% [–5·1 to 0·6], p=0·14). CA125 at 35 IU/mL had similar sensitivity (difference of –2·1% [–6·6 to 2·3], p=0·42) but higher specificity (difference of 6·7% [4·3 to 9·1], p<0·0001). In a post-hoc analysis, when compared with RMI1 at 250, ORADS achieved similar sensitivity (difference of –2·1%, 95% CI –8·6 to 4·3, p=0·60) and lower specificity (difference of 10·2%, 95% CI 6·8 to 13·6, p<0·0001).

Interpretation In view of its higher sensitivity than RMI1 at 250, despite some loss in specificity, we recommend that IOTA ADNEX at 10% should be considered as the new standard-of-care diagnostic in ovarian cancer for postmenopausal patients.

Lancet Oncol 2024; 25: 1371–86

This online publication has been corrected. The corrected version first appeared at thelancet.com/oncology on October 1, 2024

See **Comment** page 1251

*Collaborators are listed at the end of the Article

Pan Birmingham Gynaecological Cancer Centre, Sandwell and West Birmingham Hospitals NHS Trust, Birmingham, UK (Prof S Sundar MPhil MRCOG, F L Kwong MRCOG); **Institute of Cancer and Genomic Sciences** (Prof S Sundar), **Institute of Applied Health Research** (R Agarwal PhD, C Davenport FFPH PhD, K Scandrett MSc, Prof J Deeks PhD), **and Birmingham Clinical Trials Unit** (L Sturdy BSc, R Ottridge MPhil), **University of Birmingham, Birmingham, UK; University Hospital Southampton NHS Foundation Trust, Southampton, UK** (S Johnson FRCOG); **County Durham and Darlington NHS Foundation Trust, Darlington, UK** (P Sengupta FRCOG); **West Hertfordshire Hospitals NHS Trust, Watford, UK** (R Selvi-Vikram FRCOG); **Centre for Medical Imaging, University College London, London, UK** (S Mallett PhD); **School of Medicine, University of Nottingham, Nottingham, UK** (C Rick PhD); **St Peter's College, University of Oxford, Oxford, UK** (Prof S Kehoe FRCOG MD); **Department of Development and Regeneration** (Prof D Timmerman MD PhD, Prof B Van Calster PhD) **and Leuven Unit for Health Technology Assessment Research (LUHTAR)** (Prof B Van Calster), **KU Leuven, Leuven, Belgium; Department of Obstetrics and Gynecology, University Hospitals KU**

Leuven, Leuven, Belgium (Prof D Timmerman); Faculty of Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK (Prof T Bourne FRCOG PhD); ROCkeTS Project Management Group, Birmingham, UK (H Stobart MSc); University of Exeter Medical School, University of Exeter, Exeter, UK (Prof R D Neal FRCGP PhD); Department of Women's Cancer, Elizabeth Garrett Anderson Institute for Women's Health, University College London, London, UK (Prof U Menon MD PhD, A Gentry-Maharaj PhD); MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK (Prof U Menon, A Gentry-Maharaj); NIHR Birmingham Biomedical Research Centre, University Hospitals Birmingham NHS Foundation Trust, University of Birmingham, Birmingham, UK (Prof J Deeks, K Scandrett)

Correspondence to: Prof Sudha Sundar, Pan Birmingham Gynaecological Cancer Centre, Edgbaston, Birmingham, B15 2TT, UK
s.s.sundar@bham.ac.uk

See Online for appendix

Funding UK National Institute for Health and Care Research.

Introduction

The global incidence of ovarian cancer is estimated to be around 310 000 people per year, with a mortality rate of more than 200 000 deaths per year.[1] Unfortunately, most patients with ovarian cancer will be diagnosed at advanced stages, and the 10-year survival rate has remained static over the past decade in high-income countries, at around 35%.[2] An earlier, more accurate method of diagnosing ovarian cancer could therefore improve survival.

Ovarian cancer is associated with non-specific symptoms of persistent abdominal distension (eg, bloating); feeling full, loss of appetite, or both; pelvic or abdominal pain; increased urinary urgency, frequency, or both; unexplained weight loss; fatigue; or changes in bowel habits. Most patients referred with these symptoms and abnormal test results will not have ovarian cancer; only about 3% of premenopausal and 18% of postmenopausal individuals referred through rapid-access referrals (an expedited referral pathway) by the UK's National Health Service (NHS) will be diagnosed with ovarian cancer.[3] General practitioners and family physicians are encouraged to obtain a detailed history and examine such patients before testing those with symptoms for the CA125 tumour marker and doing a pelvic ultrasound.[4,5] In the UK, patients with abnormal test results, either for CA125 or on ultrasound, are referred to a hospital for assessment by gynaecologists via rapid-access referrals. Hospital gynaecologists then use risk-prediction models, tests, or scores to triage patients to tertiary care for specialist surgical management for gynaecological cancer. Accurate triage with rapid referral by both primary care practitioners and hospitals is important because patients with ovarian cancer who are managed with maximal cytoreduction surgery in specialist gynaecological cancer centres have better survival than those who have less extensive surgery, and because it concentrates cancer-care resources for those most at risk.[6]

Multiple risk-prediction models combining clinical, biomarker, and ultrasound indicators are used in

Research in context

Evidence before this study

We searched OVID MEDLINE, OVID Embase, and the Cochrane Library for articles published from database inception to June 3, 2024, using the search terms "ROMA", "IOTA ADNEX", "ORADS", "IOTA simple rules", and "RMI". We did not find any head-to-head prospective study that compared all of these tests in a given or predefined patient population against the same reference standard. We found several studies investigating different combinations of these tests. Studies had been mostly conducted in high-prevalence populations (ie, >35% of participants went on to be diagnosed with ovarian cancer, often advanced stage) and in specialist hospital settings with ultrasound undertaken by experts. These features reduce the applicability of the findings of these existing studies in non-specialist hospitals or community practice settings, in which triage tools are most utilised.

Added value of this study

The Refining Ovarian Cancer Test Accuracy Scores (ROCkeTS) study has identified the best diagnostic test to triage postmenopausal patients presenting in real-life clinical practice with symptoms of ovarian cancer and abnormal test results, by investigating all commonly used clinical risk-prediction models in a head-to-head prospective, high-quality, diagnostic accuracy study using a common reference standard of histology or follow-up in a predefined patient population with clear inclusion and exclusion criteria, reducing the potential for confounding and increasing the validity of the test comparisons. Since ROCkeTS recruited only newly presenting patients with symptoms, mainly recruiting participants at their first presentation to hospital (rapid-access clinics), the ROCkeTS population had a lower prevalence of ovarian cancer (17%), had more early-stage cancers (42%), and were more applicable for the evaluation of risk-prediction models than populations from previously published studies.

Implications of all the available evidence

The results of ROCkETS show a very high sensitivity for the International Ovarian Tumour Analysis (IOTA) Assessment of Different Neoplasias in the Adnexa (ADNEX) ultrasound-based model. The ROCkeTS study also showed that a high accuracy can be achieved with ultrasound-based risk-prediction models performed by non-expert sonographers, which is valuable for clinical practice. The IOTA ADNEX model at a threshold of 10% is likely to significantly improve the sensitivity of ovarian cancer risk prediction, and we recommend that it should replace the standard-of-care diagnostic test (Risk of Malignancy Index 1) for postmenopausal patients in the UK. Implementation into clinical practice is likely to increase false positives and will need to be carefully monitored by introducing additional complex imaging to decrease the burden of false positives to individuals and health systems. ROCkeTS reinforces the need for risk-prediction models to be prospectively evaluated in high-quality clinical trials in relevant populations before endorsement in guidelines and implementation in practice. The performance characteristics of the Ovarian Adnexal and Reporting Data System (ORADS) need further investigation in prospective studies. Future research will need to investigate how rapidly developing novel technologies, such as artificial intelligence, can be integrated alongside these validated models.

practice for triage in hospitals globally, including the Risk of Malignancy Index 1 (RMI1; current standard of care in the NHS), the Risk of Malignancy Algorithm (ROMA), the Assessment of Different Neoplasias in the Adnexa (ADNEX) specialist ultrasound model devised by the International Ovarian Tumour Analysis consortium (IOTA), and the Ovarian Adnexal and Reporting Data System (ORADS) ultrasound model devised by the American Radiology Association, which was introduced into clinical practice in 2020, but has not yet been prospectively validated.[7–10] In the UK, patients with an RMI1 score of greater than 250 are triaged to tertiary cancer centres for further management by gynaecological oncology surgeons, whereas patients with an RMI1 score of lower than 250 are managed in secondary care by gynaecologists.

The data underpinning these recommendations are derived from studies that predominantly include a high proportion of patients with cancer, mostly at advanced stages, and highly preselected patients who have been referred to cancer centres, making it unclear as to whether these risk-prediction models perform well when used in real-world settings of lower cancer prevalence.

A Cochrane systematic review investigating risk-prediction models for ovarian cancer included 58 studies, mostly conducted in high prevalence (ie, >35% of patients referred had cancer), specialist hospital settings with ultrasound conducted by experts.[11] Most studies were characterised by populations with a high proportion of advanced-stage cancers, in which clinical suspicion of ascites and peritoneal disease is likely to trigger CT imaging and biopsy as first steps, making triage with minimally invasive prediction models irrelevant. These features limit the applicability of risk-prediction models to non-specialist hospitals or community practices where triage tools are most used. Moreover, a systematic review by Bossuyt and colleagues[12] highlights the poor quality of diagnostic accuracy studies in ovarian cancer with the majority showing spin (ie, misrepresentation and overinterpretation that results in unjustified optimism in the study results about the performance of putative biomarkers).

For a risk-prediction model to be clinically relevant, it needs to have high diagnostic accuracy in low-prevalence settings to discriminate early-stage cancer from benign histology, ascertained in newly presenting populations. Ultrasound interpretation is influenced by practitioner expertise, and therefore ultrasound models need to be evaluated when performed by non-expert practitioners, who undertake most scans. Furthermore, model performance needs to be reported separately in premenopausal and postmenopausal populations because the prevalence of ovarian cancer and predominant histology type differ between these groups.[3,11]

In the Refining Ovarian Cancer Test Accuracy Scores (ROCkeTS) study, we investigated the accuracy of risk-prediction models for diagnosing ovarian cancer in newly presenting, symptomatic premenopausal and postmenopausal patients, with ultrasound models performed by non-experts.[13] In this Article, we present the results for the postmenopausal cohort.

Methods

Study design and participants

In this multicentre, prospective diagnostic accuracy study, newly presenting female patients aged 16–90 years who had been referred to hospital with non-specific symptoms as described by the UK National Institute for Health and Care Excellence guidance[5] and raised CA125 values or abnormal ultrasound findings, as interpreted by the primary care practitioner, were prospectively and consecutively recruited from 23 hospitals in the UK (appendix pp 1, 34). We began recruiting patients at outpatient clinics (rapid-access referrals, ultrasound clinics, routine primary care referrals, or cross-specialty referrals) or as inpatients through emergency presentations to secondary care. Exclusion criteria were pregnancy, declining a transvaginal scan, active non-ovarian malignancy, or previous ovarian malignancy. From March 14, 2018, a protocol amendment approved by the NHS West Midlands Research and Ethics Committee also excluded patients with a simple ovarian cyst of less than 5 cm in diameter and normal CA125 concentrations, because these patients have a very low risk of malignancy. Sex and ethnicity were self-reported by the participant to study staff. We did not collect information on how sex and ethnicity were defined in the study. Recruitment was conducted by research nurses and delivered through the National Collaborative Research Network (appendix pp 31–33).

Because most risk-prediction models either have different thresholds or incorporate different covariates according to menopausal status, all participants, including perimenopausal participants, were classified into dichotomous groups of premenopausal or postmenopausal on the basis of a patient-expressed history of vaginal bleeding to enable accurate analysis, with those who had not had a period for more than 12 months classified as postmenopausal. All other participants were classified as premenopausal and were excluded from the analyses reported in this Article.

ROCkeTS received ethical approval from NHS West Midlands Research and Ethics Committee (14/WM/1241) and is registered with the International Standard Randomised Controlled Trial Number registry (ISRCTN 17160843). The trial protocol has been previously published[13] and is available online. Written informed consent was obtained from participants before participation. Our report adheres to the STARD and TRIPOD reporting guidelines (appendix pp 22–25).[14,15]

For the **protocol** see https://www.birmingham.ac.uk/research/bctu/trials/pd/rockets

Procedures

All participants completed a symptom questionnaire, gave a blood sample, and had transabdominal and

transvaginal ultrasound scans at recruitment, which formed the components of the index tests. Data on age, height, weight, race, smoking status, alcohol consumption, and medical history were also obtained via a participant questionnaire done at baseline.

The index tests assessed in this study are briefly described here; more detail can be found in the appendix (pp 25–29). The ROMA risk-prediction model combines measurements of CA125 and HE4 tumour markers with menopausal status to obtain a risk probability of an ovarian cancer diagnosis via a blood test; the manufacturer (Roche Diagnostics, Welwyn Garden City, UK) recommends using a threshold score of 29·9 to trigger referral to a tertiary care centre for postmenopausal patients with a pelvic mass, but threshold scores of 14·4, 25·3, and 27·7 have been used in previous studies and were therefore also assessed in our study.[11] Ultrasound models and scores that were evaluated were the IOTA ADNEX risk-prediction model (thresholds of 10% and 3%), IOTA Simple Rules (malignant *vs* benign, or inconclusive), and the IOTA Simple Rules Risk (SRRisk) model (thresholds of 10% and 3%).[9,16–19] IOTA Simple Rules and the IOTA ADNEX and IOTA SRRisk models were developed and validated using the following: primary invasive ovarian cancer, secondary malignancy, or borderline tumours. ROMA was developed and validated using primary invasive ovarian cancer alone. CA125 concentration at a threshold of 35 IU/mL was also evaluated via blood test.

The comparator test was the RMI1 risk-prediction model, which combines measurements of CA125 and some ultrasound features, and has a threshold of 250 points. We also investigated RMI1 at a threshold of 200 as an index test.

In 2020, after completion of ROCkeTS recruitment, the ORADS scoring system was devised by the American Radiology Association, based on a set of expert consensus-agreed variables from IOTA study data.[9] The ORADS ultrasound system uses lexicon terminology for describing imaging characteristics of lesions and uses six risk assessment categories (ORADS 0–5) to describe low risk (ORADS lexicon 1) to high risk (ORADS lexicon 5) of malignancy. At the time of reporting, ORADS has not yet been prospectively validated. In post-hoc analyses, we mapped IOTA variables from the ROCkeTS ultrasound case-report forms onto the ORADS lexicon using methods described previously to calculate ORADS scores, using a threshold of 10% (equivalent to ORADS lexicon 4–5 [intermediate to high risk]).[20]

Serum samples were collected as per a predefined standard operating procedure, transported, and stored at –80°C until analysis at NHS South Tyne and Wear Clinical Pathology Services laboratory (Queen Elizabeth Hospital, Gateshead, UK). For analysis, samples were thawed in batches. Testing for HE4 and CA125 was performed on Roche Cobas e802 (Roche Diagnostics, Indianapolis, IN, USA) as per manufacturer recommendations. CA125 and HE4 were measured with electrochemiluminescence immunoassay technology (Roche Elecsys assay kits, Roche Diagnostics) according to the manufacturer's instructions.

Sonographers at participating sites received 1-day in-person and online ultrasound training and were then assessed by a written examination conducted by the IOTA team. Ultrasound within the ROCkeTS study was permitted to be conducted only by those who passed IOTA certification. Additionally, a sample of ultrasound images and reports were centrally reviewed by the IOTA team of ultrasound experts in a quality assessment to check that imaging was annotated as per IOTA terminology. Scans were performed mainly by level 2 ultrasound examiners (ie, non-medical sonographers). However, no minimum experience was stipulated for sonographers to be able to participate in ROCkeTS. Ultrasound examiners passed the quality assessment if their first three scans were accurately annotated (ie, seven of eight features were accurate); if not, the first ten scans were reviewed. Ultrasound examiners who failed the quality assessment received feedback and resubmitted images for quality assessment review after reviewing online IOTA resources. The emphasis within ROCkeTS was the evaluation of risk-prediction models; therefore, ultrasound examiners who had assessed the lesion correctly on subjective assessment but who had not annotated the image accurately were deemed to have failed quality assessment. Ultrasound examiners who completed fewer than ten scans for ROCkeTS were not assessed. Those who failed or did not undergo quality assessment were allowed to participate in the study, but a secondary analysis was done including only data from those assessed by sonographers who passed the quality assessment.

All risk-prediction tests were conducted within 3 months of participant recruitment and presentation, and before surgery or biopsy (if appropriate). Those assessing the results of the index tests were masked to the results of the reference standard. The reference standard was histology or cytology from surgery or biopsy. Pathology data were derived from pathology reports by specialist gynaecological pathologists. For participants who did not undergo surgery or biopsy for a reference standard, any subsequent diagnosis of cancer or any other medical condition was ascertained with a questionnaire completed by the participant and the research nurses at 12 months after study recruitment (appendix pp 35–40). We did not stipulate a follow-up protocol for participants within the study; participants who did not have surgery or a biopsy within 3 months of study recruitment were managed as per local protocols.

Participants were removed from the study if they withdrew consent after recruitment. Patients could opt for partial withdrawal (in which case clinical data collected up to the point of withdrawal could be used), or for complete withdrawal (in which case no data could be

used). Participants could also be withdrawn by investigator decision if deemed ineligible after recruitment; these patients would not be followed up.

Safety was assessed continuously throughout the study. As there are no foreseeable risks of mortality or substantial morbidity associated with testing, only serious adverse events believed to be associated with any study procedures were reported. The collection and reporting of serious adverse events was in accordance with Good Clinical Practice and the Research Governance Framework 2005.

Outcomes

The primary outcome was the diagnostic accuracy of index tests for diagnosing ovarian cancer (binary outcome), defined as primary invasive malignant ovarian neoplasms (versus benign or normal histology), as confirmed by histology from surgery or biopsy or at the 12-month follow-up. Primary invasive ovarian cancer was defined as cancer in the ovaries or fallopian tube, or primary peritoneal cancer. Diagnostic accuracy was assessed by sensitivity, specificity, and positive predictive value (PPV) or negative predictive value (NPV) at different thresholds. Model performance was further assessed in terms of discrimination (*C*-index) and calibration (observed *vs* predicted probabilities). We did not choose a single measure of accuracy as a primary endpoint a priori. This approach was chosen to fully evaluate the trade-offs inherent in the performance of diagnostic tests.

The main secondary outcome was the diagnostic accuracy of the index tests for diagnosing ovarian cancer (binary outcome) defined as either primary invasive ovarian cancer or secondary malignant, borderline neoplasms or neoplasms of uncertain or unknown behaviour (versus benign or normal histology), as confirmed by histology of surgical or biopsy samples or cytology alone or at the 12 month follow-up. A prespecified analysis of this secondary outcome was also performed by grouping participants with borderline neoplasms into the benign or normal histology category. To understand variability in test performance, particularly for ultrasound models, other prespecified secondary outcomes were diagnostic accuracy of the index tests for diagnosing the secondary outcome definition of ovarian cancer in the subset of participants in whom ultrasound scans were performed by examiners who passed the IOTA quality assessment, and in the subset of participants assessed in one-stop clinics (ie, centres offering ultrasound during the same appointment as the gynaecological consultation, the results of which will be reported separately). A prespecified exploratory analysis also investigated diagnostic accuracy as per the secondary outcome definition of ovarian cancer in the subset of participants recruited in high-volume centres (ie, recruiting ≥50 participants to the study). We did not investigate interobserver variability at the individual sonographer level. A full list of all exploratory outcomes can be found in the protocol.

Statistical analysis

The original sample size was based on the performance of RMI1, which is assumed to have a sensitivity of 70% and specificity of 90%, and was calculated to provide enough participants to detect a 10 percentage point increase in sensitivity (to 80%) and a 5 percentage point increase in specificity (to 95%). Based on a prevalence of 30% of ovarian cancer in referred patients (local audit), a sample size of 1333 patients in the postmenopausal cohort would provide 90% power to detect an increase in sensitivity to 80% and in specificity to 95% in paired data (conservatively assuming independence of test errors). A review of the early data in 2016 revealed a lower-than-expected ovarian cancer prevalence of 8%. Furthermore, our systematic review on the sensitivity of all the included models suggested that sensitivity could increase to 85% (ie, a 15 percentage point difference).[11] Therefore, a sample modification was required due to the low prevalence of ovarian cancer, the assumed difference

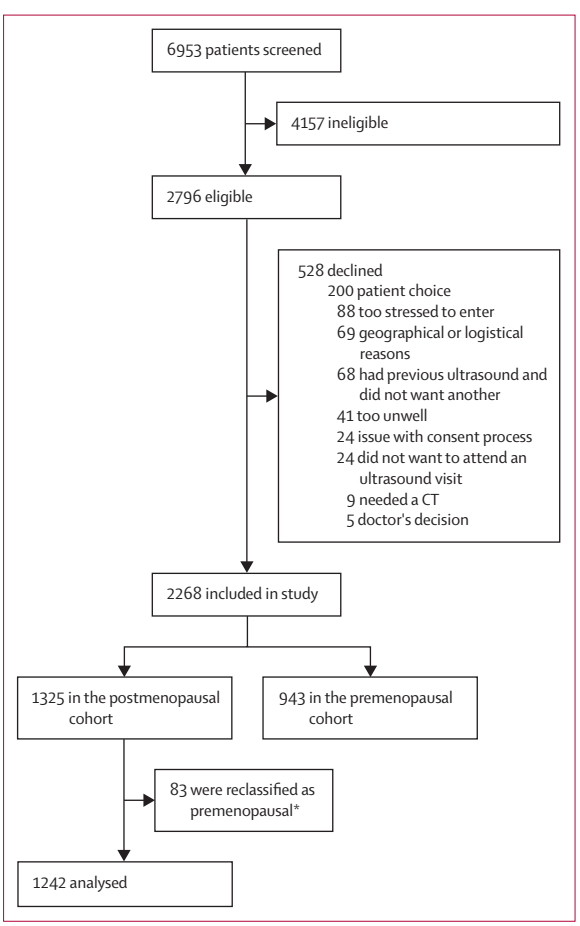

Figure 1: **Study flow diagram**
*After the study closed for recruitment, some patients were reclassified for menopausal status after analysis of their symptoms.

in sensitivity in RMI1 versus other index tests, and test error correlation because many components of the alternative tests (to be compared with RMI1) contain aspects of the RMI1, making positive test error correlation probable. Thus, the study sample size was reset based on the requirement that 150 participants with ovarian cancer would be needed to detect a 13 percentage point difference in sensitivity from 70% to 83% with 90% power, assuming a positive correlation of test error. Prevalence was monitored to ensure that the target recruitment of 150 participants with ovarian cancer was reached before study recruitment was paused.

Participants with missing or inconclusive reference standard results, or those with reference standard results that were not included within the primary outcome definition of ovarian cancer, were recorded but excluded from the primary analysis, and those with missing index test results were also excluded from the primary analysis of that particular index test. The secondary outcome definition of ovarian cancer was designed to include as many participants as possible, since it included patients with non-ovarian cancer metastatic to the ovary (secondary malignancy), borderline tumours, and tumours of unknown malignant potential as well as expanding the reference standard to include cytology in addition to histology, therefore only those with missing or inconclusive reference standard results and those with missing index test results were excluded from the secondary analysis.

Sensitivities, specificities, the *C*-index (area under the curve), and the PPV and NPV of RMI1 (threshold 200), ROMA, IOTA ADNEX, IOTA Simple Rules, the IOTA SRRisk model, and CA125 were calculated and compared with the existing RMI1 model at a threshold of 250, accounting for multiple testing with Bonferroni correction (11 pairwise comparisons, using p=0·0045 to indicate a statistically significant result). For the IOTA ADNEX and SRRisk models, we classed the 10% threshold as the primary threshold in our statistical analysis plan, and 3% as the secondary threshold. A receiver-operating characteristic (ROC) plot was created including each index test (excluding IOTA Simple Rules, which has no positvity threshold) with labels for the respective thresholds. The difference in sensitivity and specificity (and their corresponding 95% CIs) between each index test and RMI1 (250 threshold) was assessed with McNemar's test. p values were calculated with the exact McNemar test and are reported for the differences in sensitivities and specificities. The corresponding 95% CIs are exact binomial (asymptotic). Multiple testing was accounted for by use of the Bonferroni correction.[21,22]

For the risk-prediction models ROMA, IOTA ADNEX, and IOTA SRRisk, we compared the observed outcome from histology or at 12-month follow-up with the predicted risk by creating calibration plots and assessing the calibration slope. A calibration slope value of 1·0 would signify perfect agreement between the predicted probabilities and the observed probabilities. A calibration slope of less than 1·0 would indicate that a model overpredicts the risk of ovarian cancer, whereas a calibration slope of greater than 1·0 would indicate underprediction. We used the pmcalplot command in Stata version 17 to generate the calibration plots.[23] The asymptotic method was used to compute the confidence interval for the *C*-index. Post-hoc analyses investigated the diagnostic accuracy of ORADS.

A sensitivity analysis was also done for both the primary and secondary definitions of ovarian cancer that included participants with missing index test results, but excluded those with missing reference standard results. In this analysis, values for missing variables were imputed using the multiple imputation by chained equations (MICE) for predictors of index test combinations to avoid bias and make the best use of the data, by replacing missing values with plausible values based on the distribution of the observed data.[21] This method compensated for the uncertainty of the

	Diagnosis based on reference standard			All participants (N=1242)
	Ovarian cancer (n=215)	No ovarian cancer (n=861)	Other* (n=166)	
Age, years				
Median (IQR)	67·0 (59·4–73·5)	64·9 (57·4–72·7)	66·5 (58·2–74·3)	65·3 (57·9–73·3)
Missing	0	0	0	0
Height, cm				
Median (IQR)	160 (157–165)	161 (156–165)	162 (157–166)	161 (157–165)
Missing	5 (2%)	38 (4%)	4 (2%)	47 (4%)
Weight, kg				
Median (IQR)	70·5 (61·8–81·2)	71·5 (62·0–82·0)	68·8 (60·2–80·0)	70·9 (61·6–81·6)
Missing	5 (2%)	42 (5%)	5 (3%)	52 (4%)
Race or ethnicity				
White	203 (94%)	815 (95%)	151 (91%)	1169 (94%)
Asian, Bangladeshi	0	2 (<1%)	0	2 (<1%)
Asian, Chinese	1 (<1%)	3 (<1%)	0	4 (<1%)
Asian, Indian	3 (1%)	6 (1%)	3 (2%)	12 (1%)
Asian, Pakistani	0	0	0	0
Black, African	2 (1%)	1 (<1%)	0	3 (<1%)
Black, Caribbean	1 (<1%)	8 (1%)	1 (1%)	10 (1%)
Mixed	2 (1%)	4 (<1%)	0	6 (<1%)
Other	1 (<1%)	5 (1%)	0	6 (<1%)
Prefer not to say	0	3 (<1%)	1 (1%)	4 (<1%)
Missing	2 (1%)	14 (2%)	10 (6%)	26 (2%)
Smoking status				
Never	119 (55%)	451 (52%)	79 (48%)	649 (52%)
Current	18 (8%)	115 (13%)	26 (16%)	159 (13%)
Ex-smoker	76 (35%)	278 (32%)	51 (31%)	405 (33%)
Missing	2 (1%)	17 (2%)	10 (6%)	29 (2%)
Units of alcohol per week				
Median (IQR)	1·5 (0–8)	1 (0–8)	1 (0–6)	1 (0–7)
Missing	3 (1%)	26 (3%)	11 (7%)	40 (3%)

(Table 1 continues on next page)

imputation procedure and ultimately allowed us to perform the analysis on most participants, with greater power. Distributions of imputed values were visually checked for comparability with the observed data. Imputed datasets were created by replacing missing values with simulated values from a set of imputation models constructed from all predictors and the outcome variable. Multiple imputation was performed with the mi package in Stata 17. The number of imputed datasets that were created was determined by the percentage of participants who had at least one variable missing. Missing or inconclusive data for the reference standard was not imputed.

Role of the funding source

The funder of the study had no role in data collection, data analysis, data interpretation, or writing of the report, but specified the study design and choice of comparator test.

Results

Between July 13, 2015, and Nov 30, 2018, 1325 participants were recruited to the postmenopausal cohort from 23 hospitals across the UK, with follow-up ending on Nov 30, 2019 (figure 1). After the study closed for recruitment, 83 patients were reclassified for menopausal status after analysis of patient symptoms, leaving 1242 in the postmenopausal cohort.

The demographics and clinical characteristics of the 1242 postmenopausal participants are presented in table 1, stratified by the primary outcome definition of ovarian cancer versus no ovarian cancer. The median age of the participants was 65·3 years (57·9–73·3). 215 (17%) postmenopausal participants were diagnosed with the primary outcome definition of ovarian cancer; 197 (16%) were diagnosed by surgery or biopsy histology, and 18 (1%) were identified at the 12 month follow-up. The International Federation of Gynecology and Obstetrics stage of the 215 participants diagnosed with ovarian cancer was stage I in 65 (30%) patients, stage II in 25 (12%), stage III in 92 (43%), stage IV in 16 (7%), and was missing in 17 (8%). 861 (69%) participants were identified as having benign or normal histology or reported not having been diagnosed with ovarian cancer at the 12 month follow-up. Of 166 (13%) participants with missing, inconclusive, or other reference standard results, 14 (8%) had missing diagnosis data, 58 (35%) had borderline neoplasm, six (4%) had neoplasms of uncertain or unknown behaviour, ten (6%) had no histology, 22 (13%) had secondary malignant neoplasm, 20 (12%) had primary invasive malignant neoplasm in which the primary cancer site was not in the ovary or fallopian tube or was primary peritoneal (therefore considered also a secondary malignant neoplasm), nine (5%) had primary invasive malignant neoplasm for which the primary cancer site was in the ovary or fallopian tube, but the method of cancer diagnosis was cytology alone or not reported, four (2%) had a diagnostic category of "other" in the study case report form, 21 (13%) reported a diagnosis of non-ovarian cancer at the 12 month follow-up, and two (1%) had secondary cancer.

These 166 participants with missing, inconclusive, or other reference standard results were excluded from the

	Diagnosis based on reference standard			All participants (N=1242)
	Ovarian cancer (n=215)	No ovarian cancer (n=861)	Other* (n=166)	
(Continued from previous page)				
Current medical conditions				
None	52 (24%)	199 (23%)	39 (23%)	290 (23%)
Endometriosis	6 (3%)	24 (3%)	3 (2%)	33 (3%)
Adhesions	3 (1%)	14 (2%)	3 (2%)	20 (2%)
Fibroids	22 (10%)	95 (11%)	13 (8%)	130 (10%)
Adenomyosis	1 (<1%)	5 (1%)	0	6 (<1%)
Uterine polyps	13 (6%)	41 (5%)	8 (5%)	62 (5%)
High blood pressure	68 (32%)	302 (35%)	51 (31%)	421 (34%)
Epilepsy	1 (<1%)	10 (1%)	1 (1%)	12 (1%)
Heart disease	13 (6%)	69 (8%)	14 (8%)	96 (8%)
Arthritis	69 (32%)	322 (37%)	57 (34%)	448 (36%)
Uterine or bladder prolapse	14 (7%)	60 (7%)	11 (7%)	85 (7%)
Vulva pain or vulvodynia	3 (1%)	17 (2%)	4 (2%)	24 (2%)
Irritable bowel syndrome	25 (12%)	124 (14%)	25 (15%)	174 (14%)
Diverticulitis	20 (9%)	106 (12%)	16 (10%)	142 (11%)
Sexually transmitted infection	0	7 (1%)	1 (1%)	8 (1%)
High blood sugar or diabetes	26 (12%)	104 (12%)	18 (11%)	148 (12%)
Jaundice	2 (1%)	5 (1%)	2 (1%)	9 (1%)
High blood cholesterol	56 (26%)	202 (23%)	34 (20%)	292 (24%)
Pelvic inflammatory disease	0	7 (1%)	0	7 (1%)
Postmenopausal bleeding				
No	87 (40%)	407 (47%)	72 (43%)	566 (46%)
Yes	21 (10%)	124 (14%)	17 (10%)	162 (13%)
Missing	107 (50%)	330 (38%)	77 (46%)	514 (41%)
Hormonal replacement therapy				
No	149 (69%)	582 (68%)	110 (66%)	841 (68%)
Yes	6 (3%)	56 (7%)	6 (4%)	68 (5%)
Previous use	54 (25%)	196 (23%)	32 (19%)	282 (23%)
Missing	6 (3%)	27 (3%)	18 (11%)	51 (4%)
Surgical history				
None	173 (80%)	617 (72%)	119 (72%)	909 (73%)
Hysterectomy	32 (15%)	173 (20%)	34 (20%)	239 (19%)
Cystectomy	8 (4%)	60 (7%)	3 (2%)	71 (6%)
Salpingectomy	6 (3%)	30 (3%)	2 (1%)	38 (3%)
Oophorectomy	8 (4%)	39 (5%)	3 (2%)	50 (4%)

(Table 1 continues on next page)

	Diagnosis based on reference standard			All participants (N=1242)
	Ovarian cancer (n=215)	No ovarian cancer (n=861)	Other* (n=166)	
(Continued from previous page)				
Previous diagnosis of cancer				
Breast	19 (9%)	53 (6%)	14 (8%)	86 (7%)
Colon	1 (<1%)	6 (1%)	0	7 (1%)
Uterus	0	3 (<1%)	0	3 (<1%)
Cervix	3 (1%)	12 (1%)	3 (2%)	18 (1%)
Skin, non-melanoma	5 (2%)	8 (1%)	8 (5%)	21 (2%)
Skin, melanoma	5 (2%)	11 (1%)	5 (3%)	21 (2%)
Skin, melanoma status unknown	2 (1%)	3 (<1%)	1 (1%)	6 (<1%)
Lung	0	1 (<1%)	1 (1%)	2 (<1%)
Brain	0	1 (<1%)	0	1 (<1%)
Other	2 (1%)	19 (2%)	4 (2%)	25 (2%)
Family cancer history				
None	139 (65%)	564 (66%)	114 (69%)	817 (66%)
Ovary	14 (7%)	55 (6%)	3 (2%)	72 (6%)
Breast	28 (13%)	141 (16%)	21 (13%)	190 (15%)
Colon	26 (12%)	90 (10%)	15 (9%)	131 (11%)
Uterus	8 (4%)	21 (2%)	5 (3%)	34 (3%)
Sexually active				
No	124 (58%)	487 (57%)	100 (60%)	711 (57%)
Yes	62 (29%)	292 (34%)	41 (25%)	395 (32%)
Prefer not to say	26 (12%)	69 (8%)	13 (8%)	108 (9%)
Missing	3 (1%)	13 (2%)	12 (7%)	28 (2%)
Number of pregnancies				
Median (IQR)	2 (1–3)	2 (2–3)	2 (2–3)	2 (2–3)
Missing	2 (1%)	12 (1%)	11 (7%)	25 (2%)
Number of livebirths				
Median (IQR)	2 (2–3)	2 (2–3)	2 (2–3)	2 (2–3)
Missing	34 (16%)	117 (14%)	28 (17%)	179 (14%)
Number of vaginal deliveries				
Median (IQR)	2 (1–3)	2 (1–3)	2 (1–2)	2 (1–3)
Missing	33 (15%)	118 (14%)	28 (17%)	179 (14%)
Number of caesarean sections				
Median (IQR)	0	0	0	0
Missing	36 (17%)	143 (17%)	29 (17%)	208 (17%)

Data are n (%) unless otherwise specified. All participants were female. *Includes participants with missing data for a diagnosis or no histology, and those diagnosed with borderline neoplasm, neoplasms of uncertain or unknown behaviour, secondary malignant neoplasm, primary invasive malignant neoplasm for which the primary cancer site was in the ovarian or fallopian tube but the method of cancer diagnosis was cytology alone or not reported, or non-ovarian cancer identified at the 12-month follow-up, among other reasons; see the Results section of the text for further details.

Table 1: **Demographics and clinical characteristics of postmenopausal participants, by primary outcome definition of ovarian cancer**

primary analysis; therefore, data from 1076 participants were used to assess the index tests for the primary outcome definition of ovarian cancer. Table 2 provides estimates of the accuracy of RMI1, ROMA, IOTA ADNEX, IOTA SRRisk model, IOTA Simple Rules, and CA125 individually, followed by pairwise comparisons of diagnostic accuracy with the comparator test RMI1 at a threshold of 250.

RMI1 at a threshold of 250 had a sensitivity of 82·9% (95% CI 76·7–88·0) and specificity of 87·4% (84·9–89·6). Sensitivity was highest for IOTA ADNEX at a threshold of 3·0% (100·0%, 98·0–100·0), followed by ROMA at a threshold of 14·4 (97·9%, 94·7–99·4); however, the specificities of IOTA ADNEX (3·0% threshold) and ROMA (14·4 threshold) were the lowest we found, with specificities of 30·8% (27·5–34·4) for IOTA ADNEX (3·0% threshold) and 42·4% (38·9–46·0) for ROMA (14·4 threshold). All index tests generally had a high NPV, ranging from 95·6% (93·8–97·0) for RMI1 (250 threshold) to 100·0% (98·3–100·0) for IOTA ADNEX (3·0% threshold), whereas the PPV ranged from 26·8% (23·5–30·3) for the IOTA ADNEX at a threshold of 3·0% to 69·0% (61·1–76·2) for IOTA Simple Rules (table 2).

The IOTA Simple Rules was the only index test that included inconclusive results, in 226 (21%) of 1076 participants.

The *C*-index of the index tests at various thresholds ranged from 0·88 (95% CI 0·85–0·91) for the IOTA SRRisk model to 0·93 (0·91–0·95) for IOTA ADNEX (table 2). The ROC plot of the index tests (excluding IOTA Simple Rules) is shown in figure 2A, with thresholds labelled. The calibration plots and calibration slopes for ROMA, IOTA ADNEX, and IOTA SRRisk prediction models are shown in figures 2B–D. ROMA overestimated the risk for the primary outcome, since the calibration plots shows that the predicted (expected) risks are greater than the observed risk, despite the calibration slope being greater than 1 (figure 2B). On the IOTA ADNEX and SRRisk calibration plots (figure 2C, 2D), the expected risk is roughly equal to the observed risk for patients at very low risk (ie, risk <5%), but the expected risk is greater than the observed risk for patients at higher risk for the primary outcome, above about 5% risk.

Pairwise comparison of test accuracy with RMI1 at a threshold of 250, accounting for multiple testing with Bonferroni correction (11 pairwise comparisons, using p=0·0045 to indicate a statistically significant result), was available for a maximum of 980 (91%) of 1076 participants. All pairwise comparisons are shown in table 2.

28 (2%) of 1242 participants had no reference standard data, of whom 14 (50%) had missing secondary outcome data, ten (36%) had no histology, and four (14%) had a diagnostic category of "other" in the study case report form; therefore, data from 1214 (98%) of all participants were used to assess the index tests according to the secondary outcome definition of ovarian cancer. 353 (28%) of 1242 participants met the secondary outcome definition of ovarian cancer. Of these 353 participants, 206 (58%) were diagnosed with primary invasive ovarian malignant neoplasm by surgery histology, biopsy histology, or cytology, or the method was unknown, and 18 (5%) were identified at the

12-month follow-up. A further 42 (12%) participants had secondary malignant neoplasms, 58 (16%) had borderline neoplasms, and six (2%) had neoplasms of uncertain or unknown behaviour as stated on the case report form. 21 (6%) of these 353 participants reported a diagnosis of non-ovarian cancer identified at the 12 month follow-up and two (1%) were categorised as having secondary cancer from the serious adverse event form.

RMI1 at a threshold of 250 had a sensitivity of 71·2% (95% CI 65·8–76·2) and specificity of 87·4% (84·9–89·6; table 3). Sensitivity was highest for IOTA ADNEX at a threshold of 3·0%, followed by ROMA at a threshold of 14·4. However, the specificities of IOTA ADNEX (3·0% threshold) and ROMA (14·4 threshold) were the lowest we found (table 3). All index tests generally had a high NPV, ranging from 88·3% (85·8–90·5) for CA125 to 98·2% (95·5–99·5) IOTA ADNEX (3·0% threshold), whereas the PPV ranged from 37·1% (33·7–40·6) for IOTA ADNEX (3·0% threshold), to 76·4% (69·9–82·0) for IOTA Simple Rules. The *C*-index of the index tests at various thresholds ranged from 0·84 (0·81–0·87) for IOTA Simple Rules to 0·89 (0·86 to 0·91) for ADNEX. The ROC plot of index tests (excluding IOTA Simple Rules) is shown in figure 3A, with thresholds labelled. Calibration plots and slopes for ROMA, ADNEX, and IOTA SRRisk model are shown in figures 3B–D. The calibration lines are closer to reference standard lines for all three models compared with the primary analysis.

Pairwise comparison of diagnostic accuracy with RMI1 at threshold of 250, accounting for multiple testing with Bonferroni correction (11 pairwise comparisons, using p=0·0045 to indicate a statistically significant result), was available for a maximum of 1102 participants. All pairwise comparisons are shown in table 3.

In further secondary analyses, we analysed the diagnostic accuracy of the index tests according to the secondary outcome definition of ovarian cancer, but included participants with borderline tumours in the benign or typical histology category, and found that the results were consistent with the main secondary outcome analysis (appendix pp 2–5).

133 ultrasound practitioners participated in ROCkeTS, 41 of whom undertook the IOTA quality assessment;

	Diagnosis based on reference standard (n=1076)		Sensitivity (95% CI)	Specificity (95% CI)	*C*-index, AUC (95% CI)	PPV (95% CI)	NPV (95% CI)	Pairwise comparison with RMI1* (≥250 threshold)		
	Ovarian cancer (n=215)	No ovarian cancer (n=861)						N†	Difference in sensitivity (95% CI), p value	Difference in specificity (95% CI), p value
RMI1										
Available	187 (87%)	793 (92%)	..	..	0·92 (0·89–0·94)	..	..	..	..	..
Missing	28 (13%)	68 (8%)	..	..	..	..	..	..	..	..
≥200 *vs* <200	159 (74%) *vs* 28 (13%)	124 (14%) *vs* 669 (78%)	85·0% (79·1 to 89·8)	84·4% (81·6 to 86·8)	..	56·2% (50·2 to 62·0)	96·0% (94·2 to 97·3)	980	−2·1% (−4·7 to 0·5), p=0·13	3·0% (1·7 to 4·3), p<0·0001
≥250 *vs* <250 (comparator group)	155 (72%) *vs* 32 (15%)	100 (12%) *vs* 693 (80%)	82·9% (76·7 to 88·0)	87·4% (84·9 to 89·6)	..	60·8% (54·5 to 66·8)	95·6% (93·8 to 97·0)	NA	NA	NA
ROMA										
Available	191 (89%)	766 (89%)	..	..	0·92 (0·90 to 0·95)	..	..	..	..	..
Missing	24 (11%)	95 (11%)	..	..	..	..	..	..	..	..
≥14·4 *vs* <14·4	187 (87%) *vs* 4 (2%)	441 (51%) *vs* 325 (38%)	97·9% (94·7 to 99·4)	42·4% (38·9 to 46·0)	..	29·8% (26·2 to 33·5)	98·8% (96·9 to 99·7)	876	−14·3% (−20·2 to −8·4), p<0·0001	43·1% (39·3 to 46·9), p<0·0001
≥25·3 *vs* <25·3	174 (81%) *vs* 17 (8%)	207 (24%) *vs* 559 (65%)	91·1% (86·1 to 94·7)	73·0% (69·7 to 76·1)	..	45·7% (40·6 to 50·8)	97·0% (95·3 to 98·3)	876	−7·1% (−12·3 to −2·0), p=0·0042	12·3% (9·3 to 15·3), p<0·0001
≥27·7 *vs* <27·7	169 (79%) *vs* 22 (10%)	181 (21%) *vs* 585 (68%)	88·5% (83·1 to 92·6)	76·4% (73·2 to 79·3)	..	48·3% (42·9 to 53·7)	96·4% (94·6 to 97·7)	876	−4·2% (−9·5 to 1·2), p=0·14	8·6% (5·8 to 11·4), p<0·0001
≥29·9 *vs* <29·9	168 (78%) *vs* 23 (11%)	154 (18%) *vs* 612 (71%)	88·0% (82·5 to 92·2)	79·9% (76·9 to 82·7)	..	52·2% (46·6 to 57·7)	96·4% (94·6 to 97·7)	876	−3·6% (−9·1 to 1·9), p=0·24	5·2% (2·5 to 8·0), p=0·0001
IOTA ADNEX										
Available	180 (84%)	710 (82%)	..	..	0·93 (0·91 to 0·95)	..	..	..	..	..
Missing	35 (16%)	151 (18%)	..	..	..	..	..	..	..	..
≥3·0% *vs* <3·0%	180 (84%) *vs* 0	491 (57%) *vs* 219 (25%)	100·0% (98·0 to 100·0)	30·8% (27·5 to 34·4)	..	26·8% (23·5 to 30·3)	100·0% (98·3 to 100·0)	889	−17·8% (−23·9 to −11·6), p<0·0001	56·1% (52·2 to 60·1), p<0·0001
≥10·0% *vs* <10·0%	173 (80%) *vs* 7 (3%)	295 (34%) *vs* 415 (48%)	96·1% (92·2 to 98·4)	58·5% (54·7 to 62·1)	..	37·0% (32·6 to 41·5)	98·3% (96·6 to 99·3)	889	−13·9% (−20·2 to −7·6), p<0·0001	28·5% (24·7 to 32·3), p<0·0001

(Table 2 continues on next page)

	Diagnosis based on reference standard (n=1076)		Sensitivity (95% CI)	Specificity (95% CI)	C-index, AUC (95% CI)	PPV (95% CI)	NPV (95% CI)	Pairwise comparison with RMI1* (≥250 threshold)		
	Ovarian cancer (n=215)	No ovarian cancer (n=861)						N†	Difference in sensitivity (95% CI), p value	Difference in specificity (95% CI), p value
(Continued from previous page)										
IOTA SRRisk										
Available	186 (87%)	787 (91%)	..	..	0·88 (0·85 to 0·91)	..	..	..	..	..
Missing	29 (13%)	74 (9%)	..	..	..	..	..	..	..	..
≥3·0% *vs* <3·0%	172 (80%) *vs* 14 (7%)	346 (40%) *vs* 441 (51%)	92·5% (87·7 to 95·8)	56·0% (52·5 to 59·5)	..	33·2% (29·2 to 37·4)	96·9% (94·9 to 98·3)	970	−9·7% (−16·2 to −3·2), p=0·0029	31·0% (27·1 to 34·9), p<0·0001
≥10·0% *vs* <10·0%	162 (75%) *vs* 24 (11%)	230 (27%) *vs* 557 (65%)	87·1% (81·4 to 91·6)	70·8% (67·5 to 73·9)	..	41·3% (36·4 to 46·4)	95·9% (93·9 to 97·3)	970	−4·3% (−11·0 to 2·3), p=0·23	16·2% (12·6 to 19·8), p<0·0001
IOTA Simple Rules										
Available	186 (87%)	797 (93%)	..	..	0·89 (0·85 to 0·92)	..	..	..	..	..
Missing	29 (13%)	64 (7%)	..	..	..	..	..	..	..	..
Malignant *vs* benign *vs* inconclusive	107 (50%) *vs* 19 (9%) *vs* 60 (28%)	48 (6%) *vs* 583 (68%) *vs* 166 (19%)	84·9% (77·5 to 90·7)	92·4% (90·0 to 94·3)	..	69·0% (61·1 to 76·2)	96·8% (95·1 to 98·1)	755	−1·6% (−9·3 to 6·2), p=0·82	−2·2% (−5·1 to 0·6), p=0·14
CA125										
Available	214 (>99%)	860 (>99%)	..	..	0·91 (0·88 to 0·93)	..	..	..	..	..
Missing	1 (<1%)	1 (<1%)	..	..	..	..	..	..	..	..
≥35 IU/mL *vs* <35 IU/mL	184 (86%) *vs* 30 (14%)	193 (22%) *vs* 667 (77%)	86·0% (80·6 to 90·3)	77·6% (74·6 to 80·3)	..	48·8% (43·7 to 54·0)	95·7% (93·9 to 97·1)	980	−2·1% (−6·6 to 2·3), p=0·42	6·7% (4·3 to 9·1), p<0·0001

AUC=area under the curve. IOTA ADNEX=International Ovarian Tumour Analysis' Assessment of Different Neoplasias in the Adnexa. IOTA SRRisk=International Ovarian Tumour Analysis' Simple Rules Risk. NPV=negative predictive value. PPV=positive predictive value. RMI1=Risk of Malignancy Index 1. ROC=receiver operating characteristic. ROMA=Risk of Malignancy Algorithm. *Differences in sensitivities and specificities will not always equal the sensitivity or specificity in RMI1 (250) minus the corresponding value in the test being compared, since different numbers of patients were included in different index tests due to missing data; a negative value indicates a lower sensitivity or specificity for RMI1 versus the index test, and a positive value indicates a higher sensitivity or specificity for RMI1 versus the index test. †Number of participants who had available reference standard and index test data for both tests being compared.

Table 2: **Diagnostic performance statistics of combinations of index tests by primary outcome definition of ovarian cancer**

92 were not assessed as they had performed fewer than ten scans within the study. 119 (89%) of the 133 professionals conducting ultrasound within ROCkeTS were level 2 sonographers. 38 of 41 practitioners passed the quality assessment and performed scans for 1607 (71%) of 2252 patients across both the premenopausal and postmenopausal cohorts. We analysed diagnostic accuracy using data from 863 (69%) of the 1242 postmenopausal participants, for whom scans were performed by the practitioners who had passed the IOTA quality assessment, and found that the results in this subgroup were consistent with the main secondary outcome analysis (appendix pp 6–9). Finally, we analysed diagnostic accuracy in a subgroup of 840 (67%) of the 1242 participants, who had been recruited in high-volume centres, and found that the results in this subgroup were also consistent with the main secondary outcome analysis (appendix pp 10–13). Sensitivity analysis with imputation for missing index test data were consistent with the main analysis (appendix pp 14–19). Regarding the assessment of safety, no serious adverse events related to the study procedures were reported.

In post-hoc analyses, we compared results of both the primary outcome and secondary outcome for ORADS at a threshold of 10% with RMI1 at a threshold of 250 (appendix pp 20–21). For the primary outcome, ORADS had similar sensitivity to RMI1 at the 250 threshold but lower specificity. For secondary outcome, ORADS had higher sensitivity than RMI1, but lower specificity. For the primary outcome, ORADS at a 10% threshold had a sensitivity of 76·4% (95% CI 70·1–82·0) and a specificity of 78·3% (75·3–81·0). Similarly, for the secondary outcome, ORADS had a sensitivity of 73·2% (68·2 to 77·9) and a specificity of 78·3% (75·3 to 81·0; appendix pp 20–21).

Discussion

Our results show that in newly presenting symptomatic postmenopausal patients, IOTA ADNEX at the 3% and 10% thresholds, and ROMA at a threshold of 14·4 (lower than the manufacturer-recommended threshold of 29·9), have the highest sensitivity among all diagnostic tests we assessed, exceeding a sensitivity of 96%. Of these three tests, IOTA ADNEX at a threshold of 10% had the highest specificity, at 58·5%.

ROMA at a threshold of 29·9 had similar sensitivity to RMI1 at a threshold of 250, but with significantly lower specificity. Results are consistent across primary and secondary outcome analyses and in sensitivity analyses. Similar results were achieved in the subgroup of participants receiving ultrasound scans by practitioners who had passed the IOTA quality assessment and in the subgroup of participants recruited from high-volume centres. ORADS was similar to RMI1 with regard to sensitivity at diagnosing ovarian cancer according to the primary outcome defintion, but had significantly lower specificity.

Although it has a significantly reduced specificity compared with RMI1, we recommend IOTA ADNEX at 10% as the new standard of care because it has superior sensitivity and a lower drop in specificity than the other models that also achieved a sensitivity of greater than 96%. This prioritisation of sensitivity over specificity was strongly supported both by our participant and patient advocacy representatives and by policy experts in our project oversight group.

Prioritising sensitivity over specificity increases the risk of false positives, generating anxiety for patients and unnecessary workload for health systems. We have previously identified high anxiety and distress levels in women undergoing diagnostic testing for ovarian cancer; however, anxiety and distress levels are generally lower in postmenopausal women than in premenopausal women undergoing diagnostic testing for ovarian cancer.[3,24] Implementation of IOTA ADNEX into clinical care must consider mitigation of the effect of a false-positive result on the individual and health system by, for example, incorporating an MRI as an additional test for patients with a score of 10–50% on IOTA ADNEX, which would provide further evidence on whether the tumour is benign or malignant before a patient has surgery. It is important to recognise that some patients who have a false-positive result and go on to have surgery would have chosen to opt for surgery to manage symptoms of having a pelvic mass anyway, irrespective of the test result.

A health economic analysis of adopting new diagnostic standards, such as IOTA ADNEX at the 10% threshold, is underway and will offer crucial insights for health policy decision making.

Although the performance of RMI1 and IOTA Simple Rules was consistent with previous studies, the performance of ROMA, ADNEX, and ORADS differed substantially.[20,25–29] Compared with our study, the specificity of IOTA ADNEX was higher in several studies, but shows variation by centre of practice.[27] A retrospective study by Timmerman and colleagues of more than 4500 patients who had ultrasound scans performed predominantly by experts investigated the performance of an IOTA two-step strategy involving initial triage with simple descriptors followed by IOTA ADNEX, and compared it with ORADS. At the 10% risk threshold, the ORADS lexicon had a sensitivity of 92% (95% CI 87–96) and a specificity of 80% (74–85), and the IOTA two-step strategy had a sensitivity of 91% (84–95) and a specificity of 85% (80–88).[20] However, key differences between ROCkeTS and Timmerman and colleagues'

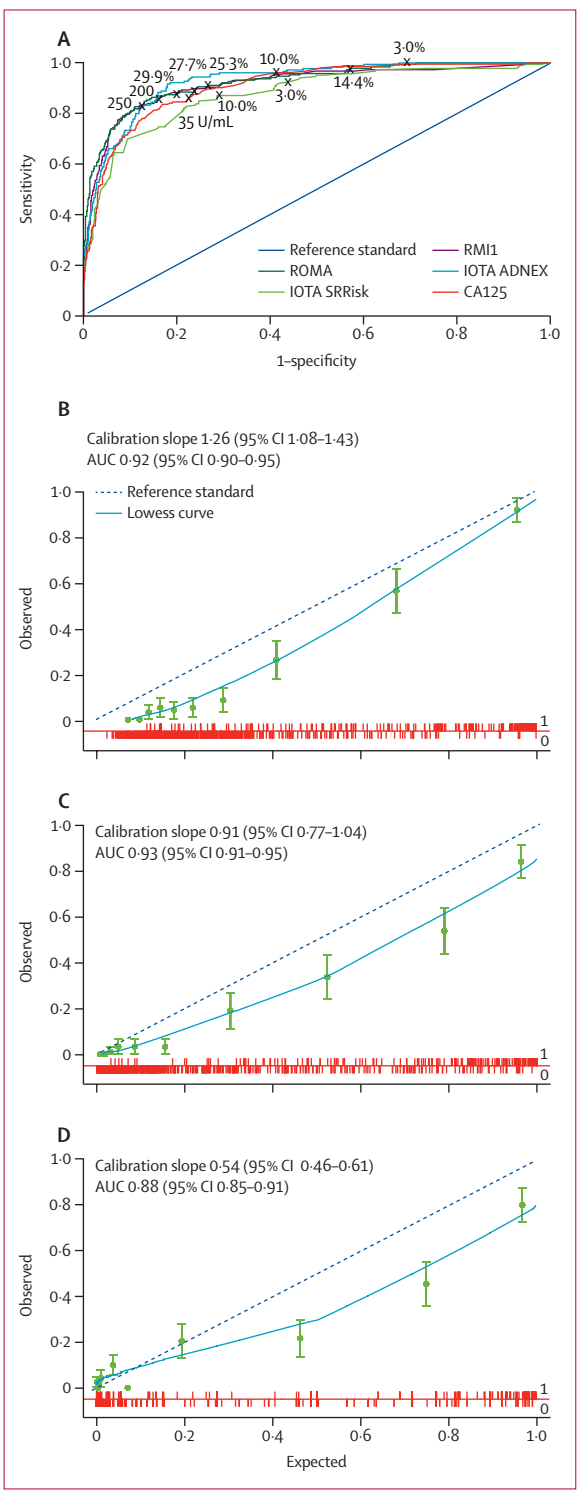

Figure 2: **ROC plot of the index tests for the primary outcome definition of ovarian cancer and calibration plots for the index tests that use a risk-prediction model**
(A) ROC plot of index test combinations. The crosses on the curves indicate the various thresholds used to indicate a participant is at risk of ovarian cancer, for each of the index tests. Calibration plots for the ROMA (B), IOTA ADNEX (C), and IOTA SRRisk (D) risk-prediction models. Calibration is shown visually by grouping participants into deciles (indicated by circles) ordered by predicted risk and considering the agreement between the mean predicted risk and the observed events in each decile. Error bars show 95% CIs. The red vertical lines in the spike plot show the distribution of participants across the risk probabilities; the 0 and 1 on the right-hand side represent whether the participants had ovarian cancer (1) or not (0). The reference standard is histology or cytology from surgery or biopsy or patient-reported diagnosis at 12 month follow-up. IOTA ADNEX=International Ovarian Tumour Analysis' Assessment of Different Neoplasias in the Adnexa. IOTA SRRisk=International Ovarian Tumour Analysis' Simple Rules Risk. RMI1=Risk of Malignancy Index 1. ROC=receiver operating characteristic. ROMA=Risk of Malignancy Algorithm.

	Diagnosis based on reference standard (n=1214)		Sensitivity (95% CI)	Specificity (95% CI)	C-index, AUC (95% CI)	PPV (95% CI)	NPV (95% CI)	Pairwise comparison with RMI1* (≥250 threshold)		
	Ovarian cancer (n=353)	No ovarian cancer (n=861)						N†	Difference in sensitivity (95% CI), p value	Difference in specificity (95% CI), p value
RMI1										
Available	309 (88%)	793 (92%)	..	..	..	..	..	..	..	..
Missing	44 (12%)	68 (8%)	..	..	0·86 (0·84 to 0·89)	..	..	..	..	..
≥200 *vs* <200	231 (65%) *vs* 78 (22%)	124 (14%) *vs* 669 (78%)	74·8% (69·5 to 79·5)	84·4 (81·6 to 86·8)	..	65·1% (59·9, 70·0)	89·6% (87·1 to 91·7)	1102	−3·6% (−5·9 to −1·2), p=0·0010	3·0% (1·7 to 4·3), p<0·0001
≥250 *vs* <250 (comparator group)	220 (62%) *vs* 89 (25%)	100 (12%) *vs* 693 (80%)	71·2 % (65·8 to 76·2)	87·4 % (84·9 to 89·6)	..	68·8% (63·4 to 73·8)	88·6% (86·2 to 90·8)	NA	NA	NA
ROMA										
Available	302 (86%)	766 (89%)	..	..	0·87 (0·85 to 0·90)	..	..	..	..	..
Missing	51 (14%)	95 (11%)	..	..	..	..	..	..	..	..
≥14·4 *vs* <14·4	280 (79%) *vs* 22 (6%)	441 (51%) *vs* 325 (38%)	92·7% (89·2 to 95·4)	42·4% (38·9 to 46·0)	..	38·8% (35·3 to 42·5)	93·7% (90·6 to 96·0)	974	−19·5% (−24·8 to −14·3), p<0·0001	43·1% (39·3 to 46·9), p<0·0001
≥25·3 *vs* <25·3	251 (71%) *vs* 51 (14%)	207 (24%) *vs* 559 (65%)	83·1% (78·4 to 87·2)	73·0% (69·7 to 76·1)	..	54·8% (50·1 to 59·4)	91·6% (89·2 to 93·7)	974	−9·0% (−13·4 to −4·6), p<0·0001	12·3% (9·3 to 15·3), p<0·0001
≥27·7 *vs* <27·7	242 (69%) *vs* 60 (17%)	181 (21%) *vs*	80·1% (75·2 to 84·5)	76·4% (73·2 to 79·3)	..	57·2% (52·3 to 62·0)	90·7% (88·2 to 92·8)	974	−5·6% (−9·9 to −1·4), p=0·0081	8·6% (5·8 to 11·4), p<0·0001
≥29·9 *vs* <29·9	238 (67%) *vs* 64 (18%)	154 (18%) *vs* 612 (71%)	78·8% (73·8 to 83·3)	79·9% (76·9 to 82·7)	..	60·7% (55·7 to 65·6)	90·5% (88·1 to 92·6)	974	−4·1% (−8·4 to 0·2), p=0·061	5·2% (2·5 to 8·0), p=0·0001
IOTA ADNEX										
Available	294 (83%)	710 (82%)	..	..	0·89 (0·86 to 0·91)	..	..	..	..	..
Missing	59 (17%)	151 (18%)	..	..	..	..	..	..	..	..
≥3·0% *vs* <3·0%	290 (82%) *vs* 4 (1%)	491 (57%) *vs* 219 (25%)	98·6% (96·6 to 99·6)	30·8% (27·5 to 34·4)	..	37·1% (33·7 to 40·6)	98·2% (95·5 to 99·5)	1003	−27·9% (−33·4 to −22·4), p<0·0001	56·1% (52·2 to 60·1), p<0·0001
≥10·0% *vs* <10·0%	270 (76%) *vs* 24 (7%)	295 (34%) *vs* 415 (48%)	91·8% (88·1 to 94·7)	58·5% (54·7 to 62·1)	..	47·8% (43·6 to 52·0)	94·5% (92·0 to 96·5)	1003	−21·1% (−26·4 to −15·8), p<0·0001	28·5% (24·7 to 32·3), p<0·0001
IOTA sRisk										
Available	306 (87%)	787 (91%)	..	..	0·85 (0·82 to 0·87)	..	..	..	..	..
Missing	47 (13%)	74 (9%)	..	..	..	..	..	..	..	..
≥3·0% *vs* <3·0%	273 (77%) *vs* 33 (9%)	346 (40%) *vs* 441 (51%)	89·2% (85·2 to 92·5)	56·0% (52·5 to 59·5)	..	44·1% (40·1 to 48·1)	93·0% (90·4 to 95·2)	1090	−17·6% (−23·2 to −12·1), p<0·0001	31·0% (27·1 to 34·9), p<0·0001
≥10·0% *vs* <10·0%	253 (72%) *vs* 53 (15%)	230 (27%) *vs* 557 (65%)	82·7% (78·0 to 86·7)	70·8% (67·5 to 73·9)	..	52·4% (47·8 to 56·9)	91·3% (88·8 to 93·4)	1090	−11·1% (−16·8 to−5·4), p=0·0001	16·2% (12·6 to 19·8), p<0·0001
IOTA Simple Rules										
Available	308 (87%)	797 (93%)	..	..	0·84 (0·81 to 0·87)	..	..	..	..	..
Missing	45 (13%)	64 (7%)	..	..	..	..	..	..	..	..
Malignant *vs* benign *vs* inconclusive	155 (44%) *vs* 51 (14%) *vs* 102 (29%)	48 (6%) *vs* 583 (68%) *vs* 166 (19%)	75·2% (68·8 to 81·0)	92·4% (90·0 to 94·3)	..	76·4% (69·9 to 82·0)	92·0% (89·6 to 94·0)	835	−5·3% (−12·0 to 1·3), p=0·13	−2·2% (−5·1 to 0·6), p=0·14
CA125										
Available	352 (>99%)	860 (>99%)	..	..	0·85% (0·82 to 0·87)	..	..	..	..	..
Missing	1 (<1%)	1 (<1%)	..	..	..	..	..	..	..	..
≥35 IU/mL *vs* <35 IU/mL	264 (75%) *vs* 88 (25%)	193 (22%) *vs* 667 (77%)	75·0% (70·1 to 79·4)	77·6% (74·6 to 80·3)	..	57·8% (53·1 to 62·3)	88·3% (85·8 to 90·5)	1102	−1·9% (−5·4 to1·5), p=0·31	6·7% (4·3 to 9·1), p<0·0001

AUC=area under the curve. NPV=negative predictive value. PPV=positive predictive value. *Differences in sensitivities and specificities will not always equal the sensitivity or specificity in RMI1 (250) minus the corresponding value in the test being compared, since different numbers of patients were included in different index tests due to missing data; a negative value indicates a lower sensitivity or specificity for RMI1 versus the index test, and a positive value indicates a higher sensitivity or specificity for RMI1 versus the index test. †Number of participants who had available reference standard and index test data for both tests being compared.

***Table 3*: Diagnostic performance statistics of combinations of index tests by secondary outcome definition of ovarian cancer**

study exist that could explain the differences in test performance. First, there were differences in the patient population; 1741 (67%) of all 2596 ROCkeTS participants (premenopausal and postmenopausal) were recruited via rapid-access referrals[3]—ie, the first point of referral to hospital (less selected)—but in Timmerman and colleagues' study, 68% of participants were recruited from cancer centres (ie, a highly preselected population).[3,30] Second, ROCkeTS was prospectively conducted with predefined inclusion and exclusion criteria, whereas Timmerman and colleagues' study was retrospective. Third, 119 (89%) of the 133 professionals conducting ultrasound within ROCkeTS were level 2 sonographers, whereas in Timmerman and colleagues' study, they were predominantly medical experts in ultrasound.

One previous study[31] investigated the performance of IOTA ADNEX in three hospitals with non-specialist sonographers, and showed that the performance of the ADNEX model was retained on external validation when conducted by ultrasound examiners with varied training and experience; however, the two participating hospitals based in the UK had previously participated in IOTA studies and were led by principal investigators with international reputations for excellence in ultrasound, and one principal investigator was an IOTA founding member. Thus, sonographers in both departments might have had access to specialist expertise not available in many NHS hospitals.[31]

Histology types and surgical outcomes from premenopausal and postmenopausal patients with ovarian cancer within ROCkeTS have been described previously,[30] and show that most participants diagnosed through symptom-triggered testing had high cytoreduction rates and a low to moderate spread of cancer. 25% of patients with high-grade serous ovarian cancer were diagnosed at stage I or stage II, reinforcing the importance of an accurate diagnosis in patients with non-specific symptoms.[30] Recruitment to ROCkeTS was predominantly through rapid-access referrals, but patients who presented as emergency admissions or elective clinic presentations were also recruited. Patients who present as emergencies are frequently too unwell to undergo full staging, which is likely to be why 8% of participants with ovarian cancer in our study had missing stage data.

We believe that the ROCkeTS study has several strengths. The study recruited only newly presenting patients with symptoms, resulting in a lower prevalence of ovarian cancer (17%), more early-stage cancers (42%), and a more applicable population for evaluation of risk-prediction models than in the previously published literature. It is a pragmatic study, reflecting the patient population that is referred from primary care or community practice to hospital, which we believe is a key strength. Our study had a relatively unselected population for assessment of diagnostic test accuracy, in contrast with many previously published studies, which had a highly pre-tested population. However, the patient population included in ROCkeTS is heterogenous with respect to the type and severity of symptoms, which reflects the conundrum faced in community and primary care.

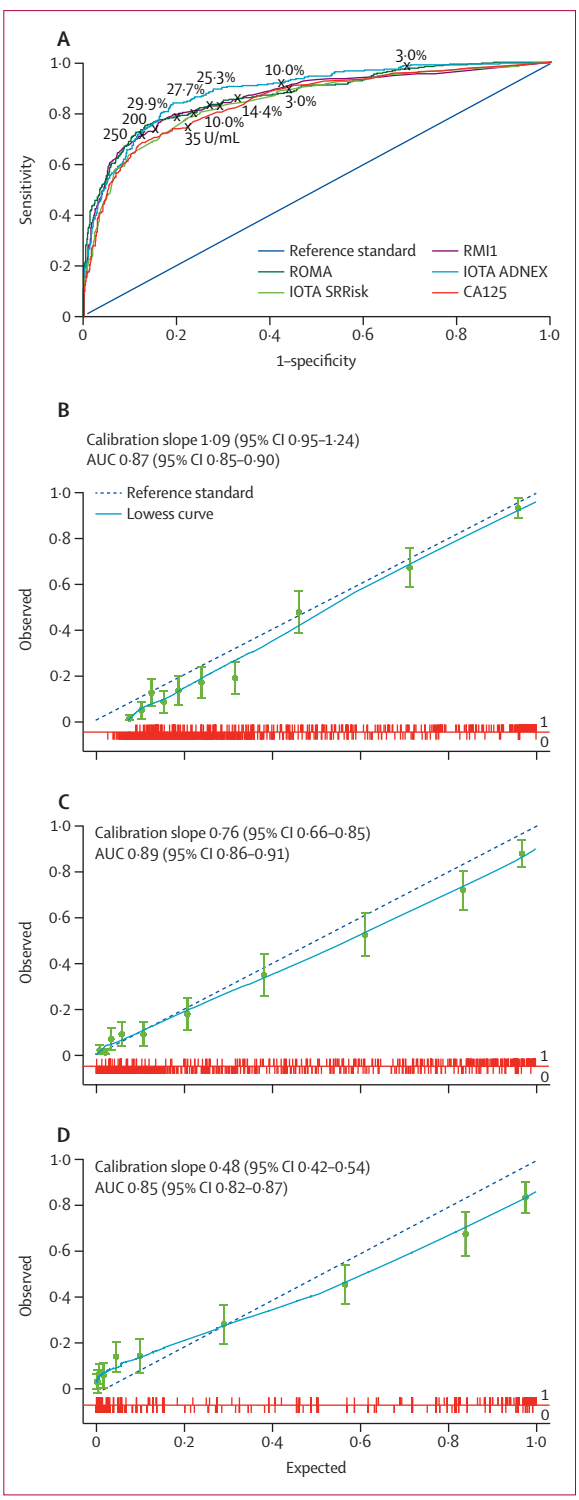

Figure 3: **ROC plot of the index test combinations for the secondary outcome definition of ovarian cancer and calibration plots for the index tests that use a risk-prediction model**
(A) ROC plot of index test combinations. The crosses on the curves indicate the various thresholds used to indicate a participant is at risk of ovarian cancer, for each of the index tests. Calibration plots for the ROMA (B), IOTA ADNEX (C), and IOTA SRRisk (D) risk-prediction models. Calibration is shown visually by grouping participants into deciles (indicated by circles) ordered by predicted risk and considering the agreement between the mean predicted risk and the observed events in each decile. Error bars show 95% CIs. The red vertical lines in the spike plot show the distribution of participants across the risk probabilities; the 0 and 1 on the right-hand side represent whether the participants had ovarian cancer (1) or not (0). The reference standard is histology or cytology from surgery or biopsy or patient-reported diagnosis at 12-month follow-up. IOTA ADNEX=International Ovarian Tumour Analysis' Assessment of Different Neoplasias in the Adnexa. IOTA SRRisk=International Ovarian Tumour Analysis' Simple Rules Risk. RMI1=Risk of Malignancy Index 1. ROC=receiver operating characteristic. ROMA=Risk of Malignancy Algorithm.

ROCkeTS was prospectively conducted with a prespecified protocol, statistical analysis plan, and sample size. Those undertaking the index tests were masked to results of the reference standard. Ultrasound training and quality assessment were mandated (although not all ultrasound practitioners did enough scans to complete the quality assessment). Recruitment was conducted by research nurses across multiple sites, reducing selection bias. Outcome data at the 12 month follow-up were ascertained robustly through information obtained directly from participants and research nurses. Missing data were appropriately handled. The statistical analysis was conducted independent of clinical investigators and ultrasound experts. We categorised patients into either premenopausal or postmenopausal groups according to patient-reported history of vaginal bleeding to address the diagnostic challenges across different stages of menopause. Results from the premenopausal cohort will be reported separately.

Moreover, our analysis carefully delineated the performance of diagnostic tests and the contribution made by metastatic ovarian cancers and borderline tumours (secondary analysis) versus that made by primary ovarian cancer alone (primary analysis).

The ROCkeTS study also has some limitations. We recruited a predominantly White population, so the results might not be as applicable to patients of other races or ethnicities. Study recruitment and follow-up was completed by October, 2019; however, analysis was delayed until 2023 due to challenges in data cleaning by sites and sample analysis in the wake of the COVID-19 pandemic. Although samples were stored at –80°C, the stability of HE4 (a key component of the ROMA test) in freeze–thaw cycles has been previously shown; therefore, the delay in analysing the samples is unlikely to have affected the results.[32] Despite the limitations of this delay, our results are still applicable for clinical practice, because ROCkeTS analysed performance of all commonly used risk-prediction models and scores used globally. Internationally, the only new diagnostic test for ovarian cancer introduced into clinical care in the past 10 years has been ORADS, which we analysed post-hoc within ROCkeTS. We followed up-to-date guidance on the interpretation and analysis of IOTA ADNEX and other risk-prediction models as recommended.[33] Although some patients with advanced-stage cancer who were too unwell or anxious did not enrol in the study, 81% of screen-eligible participants were recruited.

In real-life practice, patients undergo pelvic ultrasound delivered by sonographers with a range of experience and ROCkeTS endeavoured to replicate real-life settings as much as possible. The majority (71%) of ultrasounds within ROCKeTs (in both premenopausal and postmenopausal patients) were performed by 38 practitioners who passed ultrasound quality assessment. However, the small number of scans performed by the majority of sonographers who had not completed the quality assessment might have contributed to the lower-than-expected specificity of index tests that had an ultrasound component. However, the specificity of ultrasound-based index tests in high-volume recruiting centres was similar to the specificity across all centres combined, suggesting the specificity of index tests that had an ultrasound component reported in our study might be the true specificity in this population. We were unable to assess the contribution of two ultrasound features included in the 2024 ORADS version 2 update—bilocular cyst or shadowing for solid lesions—because these data were not collected (we used ORADS version 1); the effect of this omission is uncertain.[9,17]

Although our study assesses the performance of diagnostic tests by using accuracy measures, we have not presented data on net benefit or clinical utility, which might be as important as accuracy measures in understanding test performance, especially in the context of influencing clinical decision making.[34,35] A health economic analysis is underway and will be crucial to understand the broader effects of our findings. Moreover, it is important to note that the implications of the findings from ROCkeTS might vary across public and privately funded health systems according to the extent of guideline-compliant practice.

ROCkeTS Collaborators

Robert Kent, Natalia Rosello, Vivek Malhotra, Karen Jermy, Tim Duncan, Victoria Ames, Aarti Sharma, Anju Sinha, Majmudar Tarang, Mackenzie Ciara, Neil Hebblethwaite, Kendra Exley, Robert Macdonald, Marianne Harmer, Tracey Hughes, Rob Parker, Ahmed Darwish, Parveen Abedin, Moji Balogun, Bruce Ramsay, Roger Moshy, Mark Roberts, Michelle Russell, Ahmad Sayasneh, Ahmed Abdelbar, Shahram Abdi, Julia Palmer, Ketankumar Gajjar, Dominic Blake, Adam Naskretski, Fateh Ghazal, Harinder Rai, Patrick Keating, Nicholas Wood, Chellappah Gnanachandran, Hafez Alawad, Sonali Kaushik, Sandra Baron, Lavanya Vita, Hans Nagar, Ranjit Manchanda.

Contributors

SS, CD, SM, and JD conceptualised and designed the study. SS, SJ, PS, and RS-V recruited participants to the study with collaborators, with CR, RO, and LS coordinating the study. SM, JD, KS, FLK and RA analysed results from the study, and BVC, DT, and TB conducted the ultrasound quality assessment and training. SK, RN, RDN, UM, and AG-M provided input into the study design and conduct. HS provided a patient's perspective throughout the study, from grant application, study conduct, and interpretation of results. All authors reviewed the results and manuscript. JD, RA, and KS along with LS and RO have directly accessed and verified the underlying data reported in the manuscript. All authors had full access to all the data in the study and accept responsibility to submit for publication.

Declaration of interests

SS reports a research grant from AoA Diagnostics for work with samples collected in this study but not reported within this manuscript. SS reports honoraria from AstraZeneca, Merck, and GSK and consultancy from GSK and Immunogen, all unrelated to this work. TB reports grants, personal fees, and travel support from Samsung Medison; travel support from Roche Diagnostics; and personal fees from GE Healthcare, all outside the submitted work. BVC and DT report consultancy work done by KU Leuven to help the implementation and testing of the IOTA ADNEX model in ultrasound machines by Samsung Medison and GE Healthcare, outside the submitted work. UM declares stock ownership awarded by University College London until October, 2021, in Abcodia. UM and AG-M report research collaboration

contracts with QIMR Berghofer Medical Research Institute, iLOF (intelligent Lab on Fiber), RNA Guardian, Micronoma, Mercy Bioanalytics, and Synteny Biotechnology. SK reports an honorary role as an Ovacome charity trustee. DT, TB, and BVC are IOTA steering group members and developed the IOTA models. All other authors declare no competing interests.

Data sharing

The dataset generated, including deidentified patient data and samples analysed during the study, along with additional material such as the protocol and statistical analysis plan, will be available from the Birmingham Clinical Trials Unit, University of Birmingham, after publication. The dataset is not publicly available but may be obtained on request to SS, review by the project oversight group, UK National Institute for Health and Care Research (NIHR), ethics approval, and after fulfilling all data transfer requirements.

Acknowledgments

We thank the ROCkeTS project oversight committee (Chair—Peter Sasieni, members—Andy Nordin, Michael Weston, and Annwen Jones [Target Ovarian Cancer]) for their kind input and guidance. We acknowledge our gratitude to the patients who generously participated in our study. This study is funded by a grant from the NIHR Health Technology Assessment programme (HTA 13/13/01). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References

1 WHO. Data visualization tools for exploring the global cancer burden in 2022. https://gco.iarc.fr/today/home (accessed July 8, 2024).
2 Cancer Research UK. Ovarian cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/ovarian-cancer (accessed July 8, 2024).
3 Kwong FL, Kristunas C, Davenport C, et al. Investigating harms of testing for ovarian cancer—psychological outcomes and cancer conversion rates in women with symptoms of ovarian cancer: a cohort study embedded in the multicentre ROCkeTS prospective diagnostic study. *BJOG* 2024; **131:** 1400–10.
4 American College of Obstetricians and Gynecologists' Committee on Practice Bulletins—Gynecology. Practice Bulletin no. 174: evaluation and management of adnexal masses. *Obstet Gynecol* 2016; **128:** e210–26.
5 UK National Institute of Health and Care Excellence. Ovarian cancer: recognition and initial management. Clinical guideline [CG122]. 2011. https://www.nice.org.uk/guidance/cg122/ (accessed Sept 3, 2024).
6 UK National Institute of Health and Care Excellence. Maximal cytoreductive surgery for advanced ovarian cancer. Interventional procedures guidance [IPG757]. 2023. https://www.nice.org.uk/guidance/ipg757 (accessed Sept 3, 2024).
7 Royal College of Obstetricians and Gynaecologists. Ovarian masses in premenopausal women, management of suspected (Green-top Guideline no. 62). 2011. https://www.rcog.org.uk/guidance/browse-all-guidance/green-top-guidelines/ovarian-masses-in-premenopausal-women-management-of-suspected-green-top-guideline-no-62/ (accessed Sept 3, 2024).
8 Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. *BMJ* 2014; **349:** g5920.
9 Andreotti RF, Timmerman D, Strachowski LM, et al. O-RADS US risk stratification and management system: a consensus guideline from the ACR Ovarian-Adnexal Reporting and Data System Committee. *Radiology* 2020; **294:** 168–85.
10 Jacobs I, Oram D, Fairbanks J, Turner J, Frost C, Grudzinskas JG. 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:** 922–29.
11 Davenport C, Rai N, Sharma P, et al. Menopausal status, ultrasound and biomarker tests in combination for the diagnosis of ovarian cancer in symptomatic women. *Cochrane Database Syst Rev* 2022; **7:** CD011964.
12 Ghannad M, Olsen M, Boutron I, Bossuyt PM. A systematic review finds that spin or interpretation bias is abundant in evaluations of ovarian cancer biomarkers. *J Clin Epidemiol* 2019; **116:** 9–17.
13 Sundar S, Rick C, Dowling F, et al. Refining Ovarian Cancer Test accuracy Scores (ROCkeTS): protocol for a prospective longitudinal test accuracy study to validate new risk scores in women with symptoms of suspected ovarian cancer. *BMJ Open* 2016; **6:** e010333.
14 Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; **351:** h5527.
15 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350:** g7594.
16 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. *Gynecol Oncol* 2009; **112:** 40–46.
17 Strachowski LM, Jha P, Phillips CH, et al. O-RADS US v2022: an update from the American College of Radiology's Ovarian-Adnexal Reporting and Data System US Committee. *Radiology* 2023; **308:** e230685.
18 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* 2010; **341:** c6839.
19 Timmerman D, Van Calster B, Testa A, et al. Predicting the risk of malignancy in adnexal masses based on the Simple Rules from the International Ovarian Tumor Analysis group. *Am J Obstet Gynecol* 2016; **214:** 424–37.
20 Timmerman S, Valentin L, Ceusters J, et al. External validation of the Ovarian-Adnexal Reporting and Data System (O-RADS) lexicon and the International Ovarian Tumor Analysis 2-step strategy to stratify ovarian tumors into O-RADS risk groups. *JAMA Oncol* 2023; **9:** 225–33.
21 van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; **18:** 681–94.
22 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44:** 837–45.
23 Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024; **384:** e074819.
24 Kwong FL, Davenport C, Sundar S. Evaluating the harms of cancer testing—a systematic review of the adverse psychological correlates of testing for cancer and the effectiveness of interventions to mitigate these. *Cancers (Basel)* 2023; **15:** 3335.
25 Landolfo C, Ceusters J, Valentin L, et al. Comparison of the ADNEX and ROMA risk prediction models for the diagnosis of ovarian cancer: a multicentre external validation in patients who underwent surgery. *Br J Cancer* 2024; **130:** 934–40.
26 Van Calster B, Valentin L, Froyman W, et al. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. *BMJ* 2020; **370:** m2614.
27 Barreñada L, Ledger A, Dhiman P, et al. ADNEX risk prediction model for diagnosis of ovarian cancer: systematic review and meta-analysis of external validation studies. *BMJ Med* 2024; **3:** e000817.
28 Jha P, Gupta A, Baran TM, et al. Diagnostic performance of the Ovarian-Adnexal Reporting and Data System (O-RADS) ultrasound risk score in women in the United States. *JAMA Netw Open* 2022; **5:** e2216370.
29 Vara J, Manzour N, Chacón E, et al. Ovarian Adnexal Reporting Data System (O-RADS) for classifying adnexal masses: a systematic review and meta-analysis. *Cancers (Basel)* 2022; **14:** 3151.
30 Kwong FL, Kristunas C, Davenport C, et al. Symptom-triggered testing detects early stage and low volume resectable advanced stage ovarian cancer. *Int J Gynecol Cancer* 2024; published online Aug 13. https://doi.org/10.1136/ijgc-2024-005371.
31 Sayasneh A, Ferrara L, De Cock B, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model: a multicentre external validation study. *Br J Cancer* 2016; **115:** 542–48.

32 Sandhu N, Karlsen MA, Høgdall C, Laursen IA, Christensen IJ, Høgdall EV. Stability of HE4 and CA125 in blood samples from patients diagnosed with ovarian cancer. *Scand J Clin Lab Invest* 2014; **74:** 477–84.

33 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.

34 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352:** i6.

35 Drubay D, Van Calster B, Michiels S. Development and validation of risk prediction models. In: Piantadosi S, Meinert CL, eds. Principles and practice of clinical trials. Cham: Springer International Publishing, 2019: 1–22.