FULL PAPER

BJC

British Journal of Cancer (2016) 115, 542–548 | doi: 10.1038/bjc.2016.227

Keywords: diagnostic imaging; ovarian neoplasm; statistical models; ultrasonography

Evaluating the risk of ovarian cancer before surgery using the ADNEX model: a multicentre external validation study

A Sayasneh*[,1,2,10], L Ferrara[3,4,10], B De Cock[5], S Saso[3], M Al-Memar[3], S Johnson[6], J Kaijser[7], J Carvalho[3], R Husicka[3], A Smith[8], C Stalder[3], MC Blanco[4], G Ettore[4], B Van Calster[5], D Timmerman[5,9] and T Bourne[1,3,5]

[1]Department of Surgery and Cancer, Hammersmith Campus, Imperial College London, Du Cane Road, London W12 0HS, UK; [2]Department of Obstetrics and Gynaecology, Guy's and St Thomas' Hospital, Westminster Bridge Road, London SE1 7EH, UK; [3]Early Pregnancy and Acute Gynecology Unit, Queen Charlotte's and Chelsea Hospital, Imperial College London, Du Cane Road, London W12 0HS, UK; [4]Department of Obstetrics and Gynecology, Garibaldi Nesima Hospital, Via Palermo 636, Catania 95122, Italy; [5]KU Leuven, Department of Development and Regeneration, Herestraat 49, Box 805, Leuven 3000, Belgium; [6]Southampton University Hospitals, Princess Anne Hospital, Southampton SO16 5YA, UK; [7]Department of Obstetrics and Gynecology, Ikazia Ziekenhuis Rotterdam, Montessoriweg 1, Rotterdam 3083 AN, The Netherlands; [8]Ultrasound Scan Department, Queen Charlottes and Chelsea Hospital, Imperial College London, Du Cane Road, London W12 0HS, UK and [9]Department of Obstetrics and Gynecology, University Hospitals Leuven, Herestraat 49, Box 7003, 3000 Leuven, Belgium

Background: The International Ovarian Tumour Analysis (IOTA) group have developed the ADNEX (The Assessment of Different NEoplasias in the adneXa) model to predict the risk that an ovarian mass is benign, borderline, stage I, stages II–IV or metastatic. We aimed to externally validate the ADNEX model in the hands of examiners with varied training and experience.

Methods: This was a multicentre cross-sectional cohort study for diagnostic accuracy. Patients were recruited from three cancer centres in Europe. Patients who underwent transvaginal ultrasonography and had a histological diagnosis of surgically removed tissue were included. The diagnostic performance of the ADNEX model with and without the use of CA125 as a predictor was calculated.

Results: Data from 610 women were analysed. The overall prevalence of malignancy was 30%. The area under the receiver operator curve (AUC) for the ADNEX diagnostic performance to differentiate between benign and malignant masses was 0.937 (95% CI: 0.915–0.954) when CA125 was included, and 0.925 (95% CI: 0.902–0.943) when CA125 was excluded. The calibration plots suggest good correspondence between the total predicted risk of malignancy and the observed proportion of malignancies. The model showed good discrimination between the different subtypes.

Conclusions: The performance of the ADNEX model retains its performance on external validation in the hands of ultrasound examiners with varied training and experience.

According to the latest statistics from the National Cancer Institute in United States, 12.1 per 100 000 women developed ovarian cancer per year between 2008 and 2012, with a mortality of 7.7 per 100 000 women (Howlader *et al*, 2015). The overall 5-year survival is estimated to be ~45.6% for all stages of the disease (Howlader *et al*, 2015). However, for early localised ovarian cancers, the 5-year

*Correspondence: A Sayasneh; E-mail: a.sayasneh@imperial.ac.uk
[10]These authors contributed equally to this work.

Received 13 December 2015; revised 4 June 2016; accepted 1 July 2016; published online 2 August 2016

survival exceeds 90% (Howlader *et al*, 2015). A combination of early diagnosis and centralised management are thought to be key factors to optimise survival (Bristow *et al*, 2013, 2014; Howlader *et al*, 2015). For early diagnosis, previous trials to evaluate ovarian cancer screening have not been successful (Kobayashi *et al*, 2008; Buys *et al*, 2011). However, recently, the United Kingdom Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) showed that screening using the risk of ovarian cancer algorithm (ROCA) doubled the number of detected primary invasive epithelial ovarian or tubal cancers (iEOCs) compared with a fixed cutoff of CA125 (Menon *et al*, 2015). The researchers also reported a significant mortality reduction with annual multimodal screening (MMS) when prevalent cases were excluded. However, the effect of this mortality reduction on final ovarian cancer screening cost effectiveness requires longer-term follow-up of the study patients (Jacobs *et al*, 2015).

A further important aspect of clinical management is that an accurate diagnosis is made when a woman presents with an ovarian mass. This is essential if women with cancer are to be referred to specialist oncology services. The International Ovarian Tumour Analysis group (IOTA) have developed and validated models and rules to characterise ovarian masses as benign or malignant (Timmerman *et al*, 2005, 2010a, b; Van Holsbeke *et al*, 2012). These models and rules have also been validated in the hands of less experienced (level II) ultrasound examiners (Sayasneh *et al*, 2013a,b).

The IOTA group has developed the multiclass ADNEX (The Assessment of Different NEoplasias in the adneXa) model that can differentiate between benign tumours, borderline tumours, early-stage primary cancers, late-stage primary cancers (stages II–IV) and secondary metastatic cancers (Van Calster *et al*, 2014). The ADNEX is based on three clinical (including CA125) and six ultrasound parameters (Van Calster *et al*, 2014), and also offers risk calculation without CA125. The model was developed and temporally validated using parameters collected by experienced (or level III) ultrasound examiners, equivalent to a UK consultant level with a special interest in gynaecological ultrasonography (Education and Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB), 2006; Van Calster *et al*, 2014). This model should facilitate the management of ovarian masses more efficiently as it allows patients to be triaged to the correct management pathway, whether for conservative follow-up, surgery at a general gynaecology unit or management at high-volume specialised cancer centres. Correctly classifying the subtype of malignancy is also of critical importance as borderline ovarian tumours and early-stage ovarian cancers can be treated less aggressively, leading to the possibility of fertility preservation in younger women (Hennessy *et al*, 2009; Darai *et al*, 2013). On the other hand, metastatic ovarian cancers should be managed according to the origin of the primary cancer (Hennessy *et al*, 2009).

The primary aim of this project was to externally validate the ADNEX model. The secondary aim was to assess the performance of the model by level II examiners with varied training (nonconsultant doctors (MDs) and sonographers) (Education and Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB), 2006; Van Calster *et al*, 2014). We hypothesised that the discriminatory performance of ADNEX would be retained, that is, it would be similar to the validation performance in the original ADNEX study.

MATERIALS AND METHODS

Setting and design. This was a multicentre cross-sectional cohort study for diagnostic accuracy. Data were collected prospectively, with the purpose of developing and validating ultrasound-based prediction models from transvaginal ultrasound examinations performed by level II ultrasound examiners (nonconsultant gynaecology specialist, gynaecology trainees doctors and gynaecology sonographers) (Education and Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB), 2006; The Royal College of Radiologists (RCR) Board of the Faculty of Clinical Radiology, 2012). The ultrasound examiners were blind to the results of the reference test, that is, the final histological outcome or in the event of cancer the stage of the disease The ADNEX model was applied by a single investigator (AS) using a dedicated excel spreadsheet. Patients were recruited from three cancer centres (Queen Charlotte's Chelsea Hospital (QCCH), London, UK; Princess Ann Hospital (PAH), Southampton, UK; and Garibaldi Nesima Hospital (GNH), Catania, Italy). The study was approved as a service evaluation audit at the UK centres and as a validation study by the hospital authority at the Italian centre. The guidelines of the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) initiative were used (Collins *et al*, 2015). Patients were recruited consecutively from September 2010 to November 2014 at QCCH, from May 2012 to May 2014 at PAH and from September 2012 to February 2015 at GNH. Patients at QCCH and PAH were also recruited to the IOTA 4 study (Sayasneh *et al*, 2013a,b). Transvaginal ultrasonography was performed using the standardised approach previously published by the IOTA group (Timmerman *et al*, 2000, 2010b). Transabdominal ultrasonography was undertaken when a large mass could not be fully evaluated transvaginally (Timmerman *et al*, 2010b).

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

The exclusion criteria were (1) pregnancy, (2) patients examined by a consultant, (3) refusal of transvaginal ultrasonography, (4) cytology rather than histology as an outcome and (5) failure to undergo surgery within 120 days of the ultrasound examination. At PAH, 8 cases were included in the final analysis, although they had the ultrasound examination more than 120 days before surgery. These cases underwent a CT scan within 120 days, confirming the persistent presence of the mass.

The NHS Caldicott report guidelines were followed in all steps of data handling (Great Britain; Department of Health, 1997). At QCCH and GNH, a secure electronic data collection system was used (Astraia Software, Munich, Germany). A unique identifier was generated automatically for each patient's record. Dedicated data collection forms and excel sheets were used at PAH. Serum CA125 was measured as per clinician's discretion or clinical practice in each centre, using Abbott Architect CA125 II (Abbott Park, IL, USA) immunoassay kit at QCCH and GNH, and UniCel DxI Immunoassay System (Beckman Coulter Inc., Brea, CA, USA) Assay at PAH.

The ADNEX model. The ADNEX model contains three clinical and six ultrasound predictors: age (in years), serum CA125 level ($U\,ml^{-1}$), type of centre (oncology centres *vs* other hospitals), maximum diameter of lesion (in mm), proportion of solid tissue, more than 10 cyst locules (yes or no), number of papillary projections (0, 1, 2, 3 or >3) acoustic shadows (yes or no) and ascites (yes or no) (Van Calster *et al*, 2014). Oncology centres were defined as 'tertiary referral centres with a specific gynaecology oncology unit'. The proportion of solid tissue is obtained as the ratio of the maximum diameter of the largest solid component and

the maximum diameter of the lesion. The ADNEX model is available online and in mobile applications (www.iotagroup.org/adnexmodel/) (Van Calster *et al*, 2014). The ADNEX model can still be calculated without including the serum CA125 value. In this study we calculated the performance of the ADNEX model with and without CA125. The temporal validation of the model with CA125 in the original paper yielded an area under the receiver operator curve (AUC) of 0.943 (0.934–0.952) to discriminate benign from malignant tumours. The model without CA125 had an AUC of 0.932 (0.922–0.941). Validation AUCs between all pairs of the five categories varied between 0.71 (stage I cancer *vs* secondary metastatic cancer) and 0.99 (benign tumours *vs* late stage primary cancer). We applied the model exactly as presented in the original publication, that is, without any changes to the model formula or coefficients.

Reference tests. The reference standard was the histopathological diagnosis of the mass after surgical removal. The excised tissues underwent histological examination at the local centre. Tumours were classified according to the WHO (World Health Organisation) classification of tumours and malignant tumours were staged according to the FIGO (International Federation of Gynaecology and Obstetrics) criteria (Tavassoli *et al*, 2003; Heintz *et al*, 2006). Histological classification was performed without knowledge of the ADNEX results or clinical and ultrasound findings for the patient. The final diagnosis was categorised into five types: benign, borderline, stage I invasive, stage II–IV invasive and secondary metastatic cancer.

Statistical analysis. There were missing values for serum CA125 and for the presence of >10 cyst locules (loc10). Missing values were handled differently for serum CA125 and loc10. The number of missing values for the latter variable was small (3%), and hence these were dealt with using single stochastic imputation based on logistic regression. Missing loc10 values were predicted by a logistic regression model with Firth correction with the following predictors: age, maximum diameter of the lesion, proportion of solid tissue, number of papillations, presence of acoustic shadows, ascites, type of ovarian tumour and type of operator. The missing serum CA125 values were handled with multiple stochastic imputation using predictive mean matching regression. As the distribution of serum CA125 was heavily skewed, the log–log transformation of CA125 was used (i.e., log(log(CA125))). In this imputation model, age, maximum diameter of the lesion, proportion solid tissue, loc10, number of papillations, presence of acoustic shadows, ascites, type of ovarian tumour, hospital and operator type were used as predictors. Using this approach, the missing values were replaced by 100 plausible values, leading to 100 completed data sets. Imputed values were back transformed to the original scale. For the ADNEX model with CA125, each of the 100 completed data sets were analysed separately and their results combined using Rubin's Rules (Rubin, 1987).

External validation of the ADNEX model with and without CA125 was performed by evaluating discrimination and calibration performance. The AUC was calculated for the basic discrimination between benign and malignant tumours using the total risk of malignancy (i.e., the sum of the estimated risks of the four malignant subtypes). The 95% confidence intervals for differences in AUCs were computed based on 1000 bootstrap samples, where for each bootstrap sample the same patients were selected across the imputed data sets (Musoro *et al*, 2014). In addition, AUCs were computed for each pair of tumour types using the conditional risk method (Van Calster *et al*, 2012b). Finally, the polytomous discrimination index was calculated (Van Calster *et al*, 2012a) that estimates the average proportion of correctly classified patients by the model when presented with five patients, one with each tumour type. Sensitivity and specificity were calculated using a 1%, 5%, 10%, 15%, 20% and 30% cutoff denoting the total risk of malignancy. Calibration of the predicted probabilities was assessed through use of calibration plots that show the relation between the observed and predicted probabilities for malignant tumours. The calibration curve was estimated by using a loess smoother (Van Calster *et al*, 2016).

RESULTS

During the study period, 751 women underwent ultrasonography by level II examiners (one associate specialist in gynaecology, 12 resident gynaecology trainees and 29 sonographers) for a pelvic mass and went through the surgical management pathway. Of these, 141 women were excluded from the final analysis for the following reasons: 65 women were examined by a consultant, 26 women had no histology result (14 only cytology, 12 no cytology or histology), 24 women had surgery >120 days from the characterising ultrasound scan, 15 women were pregnant, 5 women only had a transabdominal scan, 5 women had no surgery performed (declined or were not medically fit) and finally 1 woman who had a recurrence of cervical cancer in the pelvis a few years after radical hysterectomy and underwent a bilateral salpingo-oophorectomy was excluded as the tumour was not considered adnexal. Supplementary Table 1 presents exclusions for each centre. In the final analysis, 610 women were included (Supplementary Figure 1). Of these patients, 142 (23%) had a missing CA125 level and 17 (3%) had a missing value for loc10. Supplementary Table 2 presents the numbers of missing values for each of the study centres. The prevalence of malignancy was 30% ($n=182$), with 33% for QCCH, 32% for PAH and 19% for GNH. There were 42 (7%) borderline tumours, 47 (8%) stage I primary ovarian cancers, 69 (11%) stage II–IV primary ovarian cancers and 24 (4%) secondary metastatic cancers (see Supplementary Table 3 for a breakdown per centre). The median age was 47 years with 352 (58%) premenopausal and 258 (42%) postmenopausal women. Table 1 shows descriptive statistics of the ADNEX predictors per tumour subtype. Supplementary Tables 4–6 shows descriptive statistics per centre.

The calibration plots suggest good correspondence between the total predicted risk of malignancy and the observed proportion of malignant tumours, both for the ADNEX model with and without CA125 (Figure 1).

The AUC to differentiate between benign and malignant masses was 0.937 (95% CI: 0.915–0.954) for ADNEX with CA125 and 0.925 (95% CI: 0.902–0943) for ADNEX without CA125 (Figure 2 and Table 2). The model with CA125 showed slightly better performance (AUC difference: 0.012, 95% CI: 0.006–0.020). At risk cutoffs of 1%, 10% and 30%, sensitivities were 100%, 97% and 86% for ADNEX with CA125 (Table 3). Corresponding specificities were 12%, 68% and 84%. As in the original study, centre differences were observed with centre-specific AUCs for ADNEX with CA125 that varied from 0.90 for PAH to 0.99 for GNH (Table 2). The AUC was higher for premenopausal women (0.94) than for postmenopausal women (0.90) (Table 2): 0.939 *vs* 0.899 for the model with CA125 (difference 0.04, 95% CI -0.009 to 0.084) and 0.935 *vs* 0.873 for the model without CA125 (difference 0.062, 95% CI $0.012-0.116$).

When tumours were classified into benign, borderline, stage I invasive, stages II–IV, invasive and secondary metastatic, the model showed good discrimination between the different subtypes (Table 4). For example, discrimination between benign and stage II–IV tumours was near perfect for the model with CA125 (AUC 0.99). In comparison, the model had most difficulties discriminating between borderline and stage I tumours (AUC 0.78), though its performance is still good. The model without CA125 mainly had lower AUCs for stage II $-$ IV tumours

Table 1. Descriptive information about the patients and masses included in the study according to tumour subtype

All patients	Statistic	Benign ($n=428$)	Borderline ($n=42$)	Stage I OC ($n=47$)	Stage II–IV OC ($n=69$)	Secondary metastasis ($n=24$)
Age, years	Median (IQR)	43 (31–55)	47 (30–56)	57 (48–68)	62 (53–72)	55 (49–69)
CA125, $IU\,l^{-1}$	Median (IQR)	20 (12–39)	28 (21–64)	92 (35–209)	485 (136–1083)	66 (33–129)
Max lesion diameter, mm	Median (IQR)	72 (51–95)	128 (91–174)	146 (109–180)	110 (76–140)	90 (73–135)
Presence of solid parts	*N* (%)	142 (33%)	30 (71%)	46 (98%)	69 (100%)	22 (92%)
Proportion of solid tissue, if present	Median (IQR)	0.36 (0.18–0.78)	0.37 (0.19–0.47)	0.43 (0.30–0.67)	0.59 (0.41–1.00)	1.00 (0.58–1.00)
More than 10 locules	*N* (%)	31 (7%)	14 (33%)	13 (28%)	11 (16%)	7 (29%)
Number of papillations						
0	*N* (%)	371 (87%)	26 (62%)	33 (70%)	52 (75%)	21 (88%)
1	*N* (%)	31 (7%)	6 (14%)	1 (2%)	8 (12%)	0 (0%)
2	*N* (%)	12 (3%)	2 (5%)	5 (11%)	1 (1%)	2 (8%)
3	*N* (%)	3 (1%)	2 (5%)	2 (4%)	0 (0%)	1 (4%)
>3	*N* (%)	11 (3%)	6 (14%)	6 (13%)	8 (12%)	0 (0%)
Acoustic shadows	*N* (%)	94 (22%)	0 (0%)	6 (13%)	1 (1%)	1 (4%)
Ascites	*N* (%)	6 (1%)	1 (2%)	3 (6%)	23 (33%)	7 (29%)

Abbreviations: CA125 = cancer antigen 125; IQR = interquartile range; OC = ovarian cancer.

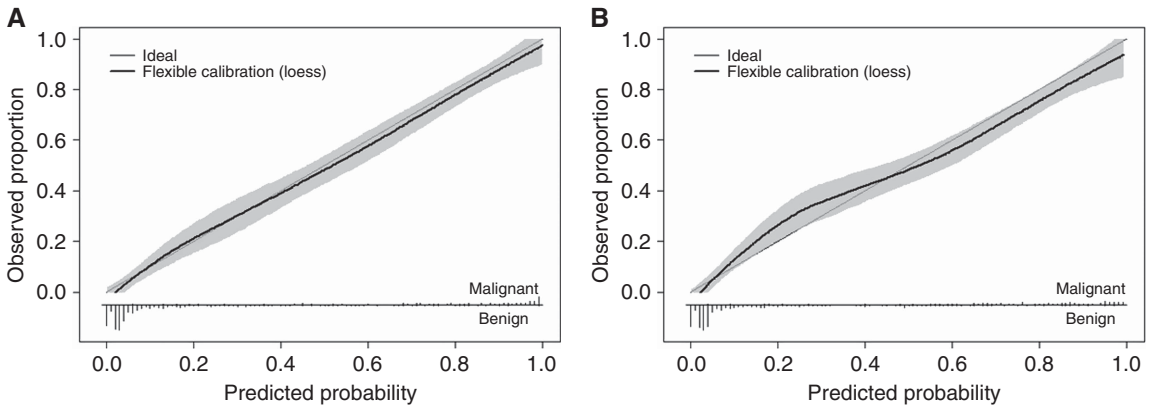

Figure 1. (**A**) Calibration plot for the ADNEX model with serum CA125. (**B**) Calibration plot for the ADNEX model without serum CA125.

vs other groups, in particular *vs* secondary metastatic cancers (AUC 0.88 for model with CA125, AUC 0.77 for model without CA125). The polytomous discrimination index (PDI) was 0.58 for ADNEX with CA125 and 0.52 for ADNEX without CA125 (Table 4), whereas PDI for random performance would be 0.20 for five categories.

DISCUSSION

In this study, we have shown that in the hands of level II ultrasound examiners, the ADNEX model was able to discriminate between benign and malignant masses with a very similar level of performance to that achieved by experienced ultrasound examiners in the original ADNEX temporal validation study published by the IOTA group (Van Calster *et al*, 2014). In our external validation study using a 10% cutoff to define malignancy, the ADNEX model achieved a sensitivity of 97.3% and a specificity of 67.7% compared with 96.5% and 71.3% in the original study (Van Calster *et al*, 2014). The optimal cutoff for selecting patients for conservative management may vary (e.g., between 1 and 5%) depending on the health-care system, cost of surgery and surgical risk factors (age, previous medical and surgical history). However, as this study only included patients who underwent surgical management, we cannot conclude which cutoff is optimal for conservative management. This will be investigated in the IOTA5 study (https://clinicaltrials.gov/ct2/show/NCT01698632). In contrast, in a tertiary centre it may be preferable to have a lower false positive rate, and a cutoff value of 30% may be more appropriate (Van Calster *et al*, 2015).

To the best of our knowledge, this is the first external validation study of the IOTA ADNEX model. Furthermore, the validation was carried out by level II ultrasound examiners, whereas in the previous IOTA development and temporal validation study (Van Calster *et al*, 2014), the ultrasound scan parameters were collected by experienced level III examiners. A strength of our study is that it is multicentre, and as it includes level II examiners with varied training and experience (sonographers and medical doctors), we think the performance of the ADNEX model in this study is likely to be generalisable. Another strength of our study is the robust selection of the reference test, as only cases with a histological outcome were included. However, this may also be seen as a weakness in relation to the potential performance of the ADNEX model for masses that are selected for conservative management as these were not included in the study. This is an issue that applies to most, if not all, of the diagnostic research carried out to date on ovarian masses. The previously mentioned IOTA 5 study should give us useful information on the diagnostic performance of ADNEX and the long-term behaviour of these masses.

A potential limitation is the use of different assay kits for serum CA125 measurements; however, the inconsistency in CA125 levels

resulting from this is thought to be limited (Davelaar *et al*, 1998). Furthermore, the variance in CA125 assay kits used in the study is a reflection of clinical reality and again means results are more likely to be reproducible (Van Calster *et al*, 2014). A further possible limitation of the study is that all three participating hospitals were referral centres for gynaecological cancers, resulting in there being a relatively high prevalence of malignant disease in the study population. Accordingly, it is possible that our findings may have limitations when trying to predict test performance either in primary care or secondary gynaecology units. However, it should be noted that in the original ADNEX study the prevalence of malignancy ranged from 0 to 66% in the 24 participating centres (Van Calster *et al*, 2014), and hence this makes it more likely that results will be generalisable. Furthermore, ADNEX explicitly corrects its prediction for type of centre (oncology centres *vs* other centres). In this sense, the potential for selection bias is accounted for by the model.

Finally, having no centralised histopathology review in our study may have led to bias. For example, distinguishing borderline tumours from benign tumours or even stage I cancer may be challenging for pathologists, where disagreement can occur and this may give inaccurate diagnostic performance results for the ADNEX model in these cases (Van Calster *et al*, 2014). However, as all the histopathology departments involved in this study were tertiary referral centres for gynaecological cancers, in the event of a discrepancy (including discrepancies in the referring units) a local review at the tertiary centre would have been held to resolve the disagreement. Furthermore, centralised review of pathology was discontinued in IOTA studies as it was shown in initial studies that there were minimal differences between local and central reports (Timmerman *et al*, 2005).

It is worth noting that we have observed variation in the ADNEX performance between centres that is comparable to the one observed in the original IOTA validation study (Van Calster *et al*, 2014). This variation could be explained by the differences in

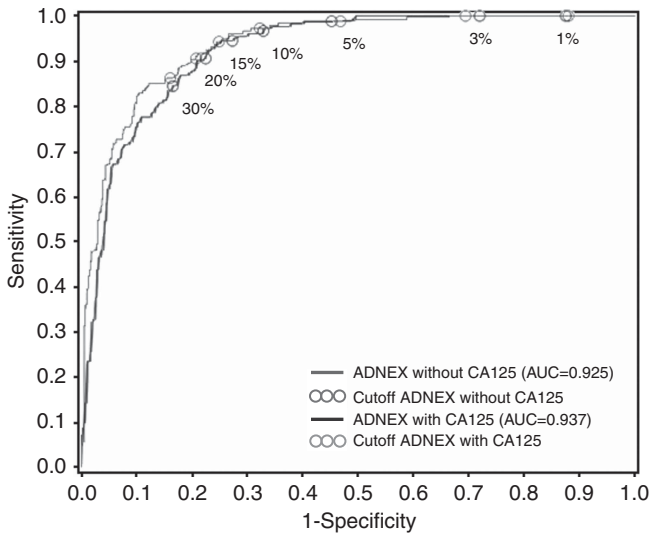

Figure 2. Receiver operating curves for the ADNEX model with and without serum CA125 levels to discriminate between benign and malignant masses.

Table 2. The area under the receiver operator curve for the discrimination between benign and malignant lesions for ADNEX with and without CA125 according to type of centre and sonographer

	ADNEX with CA125		ADNEX without CA125	
Subgroup	**AUC**	**95% CI**	**AUC**	**95% CI**
All patients	0.937	0.915–0.954	0.925	0.902–0.943
Centre				
QCCH	0.942	0.913–0.962	0.931	0.900–0.953
PAH	0.900	0.841–0.938	0.889	0.828–0.930
GNH	0.990	0.959–0.998	0.983	0.950–0.995
Operator profession				
MD	0.939	0.917–0.956	0.924	0.900–0.943
Sonographer	0.912	0.809–0.962	0.916	0.818–0.964
Menopausal status				
Premenopausal	0.939	0.901–0.963	0.935	0.901–0.958
Postmenopausal	0.899	0.855–0.931	0.873	0.824–0.910

Abbreviations: ADNEX = The Assessment of Different NEoplasias in the adneXa; AUC = area under the receiver operating curve; CA125 = cancer antigen 125; CI = confidence interval; MD = medically qualified doctor; QCCH = Queen Charlotte's and Chelsea Hospital; PAH = Princess Anne Hospital; GNH = Garibaldi Nesima Hospital.

Table 3. The overall sensitivity and specificity (benign *vs* malignant) of the ADNEX model with and without the inclusion of serum CA125

Cutoff	Patients with risk ⩾ cutoff, *N* (%)	Sensitivity with 95% CI	Specificity with 95% CI
ADNEX with CA125			
1%	559 (91.6%)	100.0% (97.4–100.0)	11.9% (9.1–15.5)
3%	479 (78.5%)	100.0% (97.4–100.0)	30.6% (26.3–35.3)
5%	383 (62.8%)	99.0% (94.9–99.8)	53.2% (48.2–58.1)
10%	315 (51.6%)	97.3% (93.5–98.9)	67.7% (63.0–72.0)
15%	281 (46.1%)	94.4% (90.0–97.0)	75.2% (70.7–79.2)
20%	253 (41.5%)	90.6% (85.2–94.1)	79.3% (75.1–83.0)
30%	226 (37.0%)	86.3% (80.4–90.6)	83.9% (80.1–87.2)
ADNEX without CA125			
1%	557 (91.3%)	100.0% (97.4–100.0)	12.4% (9.5–16.0)
3%	490 (80.3%)	100.0% (97.4–100.0)	28.0% (23.9–32.6)
5%	374 (61.3%)	98.9% (95.7–99.7)	54.7% (49.9–59.3)
10%	317 (52.0%)	96.7% (92.9–98.5)	67.1% (62.5–71.3)
15%	289 (47.4%)	94.5% (90.1–97.0)	72.7% (68.2–76.7)
20%	261 (42.8%)	90.7% (85.5–94.1)	77.6% (73.4–81.3)
30%	225 (36.9%)	84.6% (78.6–89.2)	83.4% (80.0–86.6)

Abbreviations: ADNEX = The Assessment of Different NEoplasias in the adneXa; CA125 = cancer antigen 125; CI = confidence interval. When using a 1% or 3% cutoff, confidence limits are calculated through use of Wilson's score confidence interval method with continuity correction (Newcombe, 1998). For the other cutoffs, confidence limits are calculated using logistic regression to combine results after multiple imputation.

Table 4. Pairwise AUCs and PDI of the ADNEX model with and without serum CA125

Discrimination measure	ADNEX with CA125	ADNEX without CA125
Polytomous discrimination index (PDI)	0.59	0.52
AUC benign *vs* borderline	0.88	0.88
AUC benign *vs* stage I OC	0.95	0.94
AUC benign *vs* stage II–IV OC	0.99	0.97
AUC benign *vs* secondary metastasis	0.96	0.95
AUC borderline *vs* stage I OC	0.78	0.78
AUC borderline *vs* stage II–IV OC	0.94	0.91
AUC borderline *vs* secondary metastasis	0.92	0.93
AUC stage I OC *vs* stage II–IV OC	0.83	0.79
AUC stage I OC *vs* secondary metastasis	0.81	0.83
AUC stage II–IV OC *vs* secondary metastasis	0.88	0.77

Abbreviations: ADNEX = The Assessment of Different NEoplasias in the adneXa; AUC = area under the receiver operating curve; CA125 = cancer antigen 125; OC = ovarian cancer.

the case mix between these centres with a higher number of secondary metastatic cancers in PAH compared with QCCH and GNH. It is important to investigate heterogeneity between centres, but this data set is not ideal for this objective because this requires a larger database derived from a large number of centres.

In our study, the classification of the level of experience of the ultrasound examiners (level II) was based on the recommendations published by the European Federation of Societies for Ultrasound in Medicine and Biology (Education and Practical Standards Committee, European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB), 2006) and by the Royal College of Radiologists (The Royal College of Radiologists (RCR) Board of the Faculty of Clinical Radiology, 2012). As guidance, a level III examiner in the United Kingdom equates to a consultant with a special interest in gynaecological ultrasonography (The Royal College of Radiologists (RCR) Board of the Faculty of Clinical Radiology, 2012). We acknowledge that this approach has limitations as some level II examiners may have similar levels of competence to someone with level III experience. However, it is acknowledged that the boundaries between these levels can be difficult to distinguish and may overlap (The Royal College of Radiologists (RCR) Board of the Faculty of Clinical Radiology, 2012). In our study, similar to previous findings when the IOTA model LR2 was validated in the hands of level II examiners (Sayasneh *et al*, 2013b), we found the AUC for the ADNEX model was slightly higher when the scans were performed by doctors compared with sonographers (Table 2).

By characterising the type of malignancy (borderline, primary stage I cancer, primary stage II–IV cancer or secondary metastatic), the ADNEX model offers the possibility of a more personalised diagnosis in the event of an ovarian mass. This potentially may enable fertility preserving surgery in some women, help plan the most appropriate surgical approach (laparoscopy or laparotomy) in others or direct attention to the primary site of malignancy in the event of metastasis. Although the ADNEX model gives absolute risks ratios, relative risk ratios can be computed to give a comparison with the background risk for individual patient (Van Calster *et al*, 2015). External validation is a critical step for any diagnostic test before it can be introduced into clinical practice. We have shown that the performance of the ADNEX model is retained in units with different patient populations to the original study, and that it performs well in the hands of examiners with different levels of experience and background training. Our findings suggest that the ADNEX model has the potential to improve management decisions in daily clinical practice for women with adnexal tumours.

ACKNOWLEDGEMENTS

TB 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. DT is Senior Clinical Investigator of the Research Foundation -Flanders (Belgium) (FWO). Research was supported by FWO Grants G049312N and G0B4716N and by Internal Funds KU Leuven Grant C24/15/037.

CONFLICT OF INTEREST

TB reports that clinical research in his department (QCCH, Imperial College London Healthcare NHS Trust) is supported by Samsung Medison and Roche Diagnostics. The remaining authors declare no conflict of interest.

REFERENCES

Bristow RE, Chang J, Ziogas A, Anton-Culver H (2013) Adherence to treatment guidelines for ovarian cancer as a measure of quality care. *Obstet Gynecol* **121**(6): 1226–1234.

Bristow RE, Chang J, Ziogas A, Randall LM, Anton-Culver H (2014) High-volume ovarian cancer care: survival impact and disparities in access for advanced-stage disease. *Gynecol Oncol* **132**(2): 403–410.

Buys SS, Partridge E, Black A, Johnson CC, Lamerato L, Isaacs C, Reding DJ, Greenlee RT, Yokochi LA, Kessel B, Crawford ED, Church TR, Andriole GL, Weissfeld JL, Fouad MN, Chia D, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hartge P, Pinsky PF, Zhu CS, Izmirlian G, Kramer BS, Miller AB, Xu JL, Prorok PC, Gohagan JK, Berg CD. PLCO Project Team (2011) Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial. *JAMA* **305**(22): 2295–2303.

Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Brit J Obstet Gynacol* **122**(3): 434–443.

Darai E, Fauvet R, Uzan C, Gouy S, Duvillard P, Morice P (2013) Fertility and borderline ovarian tumor: a systematic review of conservative management, risk of recurrence and alternative options. *Hum Reprod Update* **19**(2): 151–166.

Davelaar EM, van Kamp GJ, Verstraeten RA, Kenemans P (1998) Comparison of seven immunoassays for the quantification of CA 125 antigen in serum. *Clin Chem* **44**(7): 1417–1422.

Department of Health (1997) *The Caldicott Committee Report on the Review of Patient-Identifiable Information*. Department of Health: Great Britain.

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

Heintz AP, Odicino F, Maisonneuve P, Quinn MA, Benedet JL, Creasman WT, Ngan HY, Pecorelli S, Beller U (2006) Carcinoma of the ovary. FIGO 26th Annual Report on the Results of Treatment in Gynecological Cancer. *Int J Gynaecol Obstet* **95**(Suppl 1): S161–S192.

Hennessy BT, Coleman RL, Markman M (2009) Ovarian cancer. *Lancet* **374**(9698): 1371–1382.

Howlader N, Noone AM, Krapcho M, Garshell J, Miller D, Altekruse SF, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (2015) *SEER Cancer Statistics Review, 1975-2012.* Vol. 2015. National Cancer Institute: Bethesda, MD.

Jacobs IJ, Menon U, Ryan A, Gentry-Maharaj A, Burnell M, Kalsi JK, Amso NN, Apostolidou S, Benjamin E, Cruickshank D, Crump DN, Davies SK, Dawnay A, Dobbs S, Fletcher G, Ford J, Godfrey K, Gunu R, Habib M, Hallett R, Herod J, Jenkins H, Karpinskyj C, Leeson S, Lewis SJ, Liston WR, Lopes A, Mould T, Murdoch J, Oram D, Rabideau DJ, Reynolds K, Scott I, Seif MW, Sharma A, Singh N, Taylor J, Warburton F, Widschwendter M, Williamson K, Woolas R, Fallowfield L, McGuire AJ, Campbell S, Parmar M, Skates SJ (2015) Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet* **387**: 945–956.

Kobayashi H, Yamada Y, Sado T, Sakata M, Yoshida S, Kawaguchi R, Kanayama S, Shigetomi H, Haruta S, Tsuji Y, Ueda S, Kitanaka T (2008) A randomized study of screening for ovarian cancer: a multicenter study in Japan. *Int J Gynecol Cancer* **18**(3): 414–420.

Menon U, Ryan A, Kalsi J, Gentry-Maharaj A, Dawnay A, Habib M, Apostolidou S, Singh N, Benjamin E, Burnell M, Davies S, Sharma A, Gunu R, Godfrey K, Lopes A, Oram D, Herod J, Williamson K, Seif MW, Jenkins H, Mould T, Woolas R, Murdoch JB, Dobbs S, Amso NN, Leeson S, Cruickshank D, Scott I, Fallowfield L, Widschwendter M, Reynolds K, McGuire A, Campbell S, Parmar M, Skates SJ, Jacobs I (2015) Risk algorithm using serial biomarker measurements doubles the number of screen-detected cancers compared with a single-threshold rule in the United Kingdom Collaborative Trial of Ovarian Cancer Screening. *J Clinic Oncol* **33**(18): 2062–2071.

Musoro JZ, Zwinderman AH, Puhan MA, Ter Riet G, Geskus RB (2014) Validation of prediction models based on lasso regression with multiply imputed data. *BMC Med Res Methodol* **14**(1): 116.

Newcombe RG (1998) Two-sided confidence intervals for the single proportion comparison of seven methods. *Stat Med* **17**(8): 857–872.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.: Hoboken, NJ, USA.

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

Sayasneh A, Wynants L, Preisler J, Kaijser J, Johnson S, Stalder C, Husicka R, Abdallah Y, Raslan F, Drought A, Smith AA, Ghaem-Maghami S, Epstein E, Van Calster B, Timmerman D, Bourne T (2013b) Multicentre external validation of IOTA prediction models and RMI by operators with varied training. *Br J Cancer* **108**(12): 2448–2454.

Tavassoli FA, Devilee P. International Agency for Research on Cancer (2003) *Pathology and Genetics of Tumours of the Breast and Female Genital Organs*. International Agency for Research on Cancer: Lyon.

The Royal College of Radiologists (RCR) Board of the Faculty of Clinical Radiology (2012) *Ultrasound Training Recommendations for Medical and Surgical Specialties*. London. Available at https://www.rcr.ac.uk/sites/default/files/publication/BFCR(12)17_ultrasound_training.pdf (last accessed June 2015).

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

Timmerman D, Testa A, Bourne T, Ferrazzi E, Ameye L, Konstantinovic M, Van Calster B, Collins W, Vergote I, Van Huffel S, Valentin L (2005) A logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis (IOTA) group. *J Clin Oncol* **23**: 8794–8801.

Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. International Ovarian Tumor Analysis (IOTA) Group (2000) Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the International Ovarian Tumor Analysis (IOTA) Group. *Ultrasound Obstet Gynecol* **16**(5): 500–505.

Timmerman D, Van Calster B, Testa AC, Guerriero S, Fischerova D, Lissoni AA, Van Holsbeke C, Fruscio R, Czekierdowski A, Jurkovic D, Savelli L, Vergote I, Bourne T, Van Huffel S, Valentin L (2010b) Ovarian cancer prediction in adnexal masses using ultrasound-based logistic regression models: a temporal and external validation study by the IOTA group. *Ultrasound Obstet Gynecol* **36**(2): 226–234.

Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina M, Steyerberg EW (2016) A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* **74**: 167–176.

Van Calster B, Van Belle V, Vergouwe Y, Timmerman D, Van Huffel S, Steyerberg EW (2012a) Extending the c-statistic to nominal polytomous outcomes: the polytomous discrimination index. *Stat Med* **31**(23): 2610–2626.

Van Calster B, Van Hoorde K, Froyman W, Kaijser J, Wynants L, Landolfo C, Anthoulakis C, Vergote I, Bourne T, Timmerman D (2015) Practical guidance for applying the ADNEX model from the IOTA group to discriminate between different subtypes of adnexal tumors. *Facts Views Vis ObGyn* **7**(1): 32–41.

Van Calster B, Van Hoorde K, Valentin L, Testa AC, Fischerova D, Van Holsbeke C, Savelli L, Franchi D, Epstein E, Kaijser J, Van Belle V, Czekierdowski A, Guerriero S, Fruscio R, Lanzani C, Scala F, Bourne T, Timmerman D. International Ovarian Tumour Analysis Group (2014) 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* **349**: g5920.

Van Calster B, Vergouwe Y, Looman CW, Van Belle V, Timmerman D, Steyerberg EW (2012b) Assessing the discriminative ability of risk models for more than two outcome categories. *Eur J Epidemiol* **27**(10): 761–770.

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

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