A prospective evaluation of the IOTA Logistic Regression Models (LR1 and LR2) in comparison to Subjective Pattern Recognition for the diagnosis of ovarian cancer in the outpatient setting

Short title: IOTA LR versus Pattern Recognition in outpatients

Natalie Nunes[1], Gareth Ambler[2], Xulin Foo[1], Martin Widschwendter[3], Davor Jurkovic[1].

1 Gynaecological Diagnostic Outpatient Treatment Unit, University College Hospital, London, UK

2 Department of Statistical Science, University College London, UK

3 Department of Women's Cancer, University College London, Elizabeth Garrett Anderson, Institute for Women's Health, London, United Kingdom

Correspondence
Mr Davor Jurkovic GDOTU,
-1 Floor, Elizabeth Garrett Anderson
University College Hospital,
London, NW1 2BU, UK
Phone: 08451555000, Fax: 02076915861
Email: Davor.Jurkovic@nhs.net

Keywords: Ultrasound; Ovarian cancer; Adnexal tumour; Pattern recognition; IOTA, Logistic regression; LR1; LR2

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/uog.18918

Abstract

Objectives: To determine whether IOTA diagnostic models developed for pre-operative diagnosis of ovarian cancer could also be used to differentiate between benign and malignant adnexal tumours in the population of women attending gynaecology outpatient clinics.

Methods: All women referred to our outpatient clinic were first examined by a Level II ultrasound operator. In those diagnosed with adnexal tumours the IOTA LR1/2 protocol was used to evaluate the masses. The LR1 and LR2 models were then used to assess the risk of malignancy. Subsequently women were also examined by a Level 3 examiner who used pattern recognition to differentiate between benign and malignant tumours. Women with an ultrasound diagnosis of malignancy were offered surgery whilst asymptomatic women with presumed benign lesions were offered conservative management with a minimum follow-up of 12 months. The initial diagnosis was compared with two reference standards: histological findings and/or a comparative assessment of tumour morphology on follow-up ultrasound scans. All women in whom tumour classification on follow-up changed from benign to malignant were offered surgery.

Results: 489 women who had either or both of the reference standards were included into the final analysis. Their mean age was 50 years (range 16-91) and 45% of them were menopausal. 342/489 (69.9%) women had surgery and 147/489 (30.1%) were managed conservatively. The malignancy rate was 137/489 (28.0%). Overall sensitivities of LR1 and LR2 for the diagnosis of malignancy were 97.1% (95% CI: 92.7-99.2) and 94.9% (95%CI: 89.8-97.9) and specificities were 77.3% (95%CI: 72.5-81.5) and 76.7% (95%CI; 71.9-81.0) respectively ($p>0.05$). In comparison to pattern recognition [Sensitivity 94.2% (95% CI: 88.8 to 97.4); specificity 96.3% (95% CI: 93.8 to 98.0)], the specificities of IOTA models were significantly lower. ($p < 0.0001$) A significantly higher number of women would have been offered surgery for suspected cancer if women were assessed using the IOTA models instead of pattern recognition [213/489 (43.6%) versus 142/489 (29.0%)] ($p<0.001$).

Conclusions: IOTA models maintained their high sensitivity when used in the outpatient setting. Specificity was relatively low which indicates that a significant proportion of women would have been offered unnecessary surgery for suspected ovarian cancer. These findings show that IOTA models could be used as a first stage test to diagnose ovarian cancer in the outpatient setting but a different second stage test is required to minimise the number of false positive findings.

Introduction

The International Ovarian Tumour Analysis (IOTA) group developed two logistic regression models (LR1&LR2), aiming to improve the accuracy of pre-operative ultrasound diagnosis of ovarian cancer diagnosis by non-expert operators of average ability and experience. [1-3]

A number of previous original and validation studies showed that LR1 and LR2 perform well in the hands of expert operators and can facilitate pre-operative differentiation between benign and malignant tumours in women undergoing surgical treatment of adnexal tumours.[1, 4] Only a minority of women with an ultrasound diagnosis of an adnexal tumour however, require surgery and in the routine clinical practice the critical issue is not who will perform surgery but whether an intervention is required at all. There have been no studies so far, which assessed the suitability of IOTA models for the diagnosis of ovarian cancer in outpatient setting and for prioritising women with adnexal tumours for surgical interventions. The prevalence of malignancy in women attending outpatient clinics in likely to be lower, which could result in a larger number of interventions in women with benign disease even if the good diagnostic accuracy of IOTA models is maintained.

Another difficulty with using a test in the outpatient setting is that only a proportion of women would be selected for surgical treatment. Histological findings, which are traditionally used as the reference standard for assessing the accuracy of models for the diagnosis of ovarian cancer, are not applicable to population of women managed conservatively. In such circumstances

"delayed type prospective cross sectional studies" which include carefully planned and prolonged clinical follow up may provide the best evidence required to define the appropriate reference standard.[5] In the context of ovarian cancer diagnosis, it remains uncertain what is the required length of follow up and appropriate frequency of visits to determine the nature of an adnexal tumour.

In this prospective study we assessed the accuracy of IOTA LR1/LR2 models for the diagnosis of ovarian cancer in outpatient settings using histology as the first reference standard and the comparative assessment of tumour morphology on follow-up ultrasound scans as a second reference test. We also assessed the potential impact on the intervention rates of a policy, which would replace pattern recognition with the IOTA models for prioritising women with adnexal tumours for surgery

Methods

This was a single centre prospective observational study of consecutive women attending our gynaecological diagnostic unit for a variety of gynaecological complaints. We also included women who were diagnosed with ovarian cyst on previous imaging organised in primary care and were referred to us for specialist gynaecological imaging and management advice. The patients were recruited for the study between May 2009 to January 2012 and all women completed follow-up by January 2014. During the initial recruitment visit a detailed history was taken and women underwent clinical and ultrasound examinations. Women over the age of 40 who had no periods for 12 consecutive months with no other identifiable physiological, pathological or medical cause were defined as post-menopausal. Women who were over the age of 50 years and had had a hysterectomy were also classified as being postmenopausal.

All ultrasound examinations were performed by NN who was a Level II operator. She was trained in use of the IOTA protocol [1,6] but not in tumour "pattern recognition". [7-10] She was discouraged from attempting to designate a histological diagnosis or attempt to differentiate subjectively between benign and malignant tumours on ultrasound scan. Women with evidence of adnexal tumours on ultrasound scan were considered suitable for this study, but those with unilocular, anechoic cysts less than 2cm were excluded. Pregnant women and those unable to undergo a transvaginal scan were also excluded from the final data analysis.

The probability of malignancy within an adnexal mass was estimated by using the IOTA Logistic Regression Model (LR1 and LR2). Twelve variables were used for the LR1 calculation (1) personal history of ovarian cancer (yes=1, no=0): (2) current use of hormonal therapy (yes=1, no=0): (3) age of the patient (years): (4) maximum diameter of lesion (mm): (5) evidence of pain during the examination of the mass (yes=1, no=0): (6) presence of ascites (yes=1, no=0): (7) presence of blood flow within a solid papillary projection (yes=1, no=0): (8) purely solid tumour: (9) maximum diameter of the largest solid component (expressed in millimetres, but with no increase > 50 mm): (10) irregular internal cyst walls (yes=1, no=0): (11) presence of acoustic shadows (yes=1, no=0): (12) colour score (1-4 where 1 is no flow and 4 is maximum flow). The probability of malignancy was calculated using the formula $y = 1/ (1 + \exp[-z])$, where $z = – 6.7468 +1.5985 (1) – 0.9983 (2) + 0.0326 (3) + 0.00841 (4) – 0.8577 (5) + 1.5513 (6) + 1.1737 (7) + 0.9281 (8) + 0.0496 (9) + 1.1421 (10) – 2.3550 (11) + 0.4916 (12)$ as described in the original IOTA study. The probability y is dichotomised at a score of 0.1 to give a predictive diagnosis of cancer.

LR2 was calculated based on 6 of the above variables: (3), (6), (7), (9), (10) and (11). The formula to determine the probability of malignancy was $y = 1/ (1 + \exp[-z])$, where $z = - 5.3718 + 0.0354 (3) + 1.6159 (6) + 1.1768 (7) + 0.0697 (9) + 0.9586 (10) - 2.9486 (11)$ and as with LR1 the probability y is dichotomised at a score of 0.1 to give a predictive diagnosis of cancer.

The Level II operator recorded the IOTA assessments in the research file and these assessments were not available to the clinicians who made decisions about the patients' plan of care. In cases of multiple lesions, the lesions which were more likely to be malignant according to the IOTA model score, were included into the analysis, as the diagnosis of malignancy in one lesion supersedes the diagnosis of any coexisting benign lesions. Following the examination by the Level II operator, the women with adnexal tumours were re-examined independently by an expert ultrasound operator (DJ) who used subjective pattern recognition to determine the nature of the adnexal tumour. Women with suspected ovarian cancer following the expert exam were referred to our gynaecological oncology team for further management. Women with presumed benign lesions were offered choice between conservative management and surgery taking into account their clinical symptoms and personal preferences.

Women who opted for conservative management were offered regular follow up ultrasound scans starting with 6 monthly intervals for a minimum of 12 months. Women who become symptomatic during follow-up and those who requested intervention were offered surgery. In those who remained asymptomatic, follow-up ultrasound findings were compared to the initial diagnosis. Surgery was offered to all women in whom tumour classification was changed from benign to malignant on any of the follow-up ultrasound scans. Only when the data collection was completed at the end of the study, were the IOTA LR1 and LR2 calculations of the risk of malignancy performed and included in the data

collection sheets. Histopathology was the primary reference standard used. Tumours were classified according to the World Health Organisation (WHO) guidelines and malignancies were staged according to the International Federation of Gynecology and Obstetrics criteria.[11-12] For the women in whom surgery was not required, an ultrasound scan at 12 months or more after the primary scan confirming the initial diagnosis of benign lesion was used as the second reference standard.

It was determined by the local Research and Development Department that formal ethical assessment and approval was not required as the steps in the conduct of the study were routine practice in the unit. This includes morphological analysis of the tumours using the IOTA examination technique, measurement technique and terminology for the assessment of adnexal masses which are also a part of our standard clinical practice. In addition to this, therapeutic decisions were not based on the IOTA model scores.

Statistical analysis

The sample size for this study was determined using Harrell's recommendation that a validation dataset should contain at least 100 "events", that is borderline or malignant tumours.[13] Our validation dataset has 137 events, therefore it satisfies this requirement.

Initially, the data was analysed after assuming that both reference tests had perfect (100%) sensitivity and specificity. The sensitivity, specificity and overall accuracy of the three diagnostic tests (LR1, LR2 and pattern recognition) were calculated under this assumption and presented with exact 95% confidence intervals. Formal comparisons across different index tests were made using McNemar's test; the exact version of this test was used when necessary. These analyses were performed in the software package Stata 14.0 © (Stata Corp., College Station, TX, USA).

A secondary analysis was performed using a Bayesian approach similar to that of de Groot. [14] This analysis was performed to reduce the bias that may occur if the results of the alternative reference standard (follow-up ultrasound) are treated identically to the results from the preferred reference standard (histology) when the former may be of lower quality. In detail we created a probabilistic model that relates the underlying (unknown) status of the patient to the results of the corresponding reference test. This model assumes that the choice of reference test (histology or follow-up ultrasound) was related to the

underlying status of the patient (that is, patients with cancer were more likely to be assessed using histology). In addition, we assumed that the histology results were always correct whereas the ultrasound results were imperfect. Our (prior) belief was that the sensitivity and specificity of ultrasound were both probably close to 90% and almost certainly within the range 80-100%. Low information priors were used for all other parameters. OpenBUGS was used to estimate (posterior) distributions for the model parameters including patient status and the sensitivities and specificities of follow-up ultrasound and the three diagnostic tests.[15] All results are presented as medians with 95% credible intervals. To check the robustness of our results, additional analyses were run where the prior beliefs for ultrasound performance were changed to reflect worse performance.

Results

A total of 555 consecutive women attended for ultrasound assessment during the study period. 11 women were pregnant and they were excluded from the data analysis. A flow chart showing eligibility of women for the study and summary of their management is shown in Fig. 1. (Fig. 1) A total of 342/544 (62.9%) women had surgery after their first or a subsequent ultrasound scan while 147/544 (27.0%) were managed conservatively. (Table 1) 41/544 (7.5%) were lost to follow up, 13 women died soon after the diagnosis of adnexal tumour was made and one woman received chemotherapy as the primary treatment for presumed metastatic bowel cancer. Among the 13 women who

died, 5 had a non-ovarian malignancy (oesophageal (2), pancreatic, cervical and endometrial), 4 had non-malignant medical conditions (amyloidosis, chronic renal failure and bronchiectasis, dementia with urinary sepsis and alcoholic induced liver failure) and 4 had suspected ovarian malignancies. The patients ranged from 16 to 91 years of age with a mean age of 50 years. 237/544 (43.6%) of women were post-menopausal.

After excluding the 55/544 (10.1%) women who had neither reference test, the final diagnosis of a malignant tumour was made in 137/489 (28.0%) women and benign tumours were diagnosed in the remaining 352 (72.0%) women. Histological diagnosis was available in 342 women and they are shown in Table2. Majority of malignant tumours were invasive 116/137(84.7%) whilst the remaining 21/137 (15.3) were borderline. All borderline tumours were Stage I, whilst among invasive tumours) there were 26/116 (22.4%) Stage I, 5/116(4.3%) Stage II, 37/116(31.9%) Stage III and 22/116 (22.4%) were Stage IV tumours. The most common indications for surgery were suspected ovarian cancer (n=152 [44.4%]) and pelvic pain (n=75 [21.9%]). Other indications were the woman's choice (n=62 [18.1%]), part of subfertility or urological surgery (n=35 [10.2%]) and other reasons such as pressure symptoms, UKCTOCS Screening positive, prophylactic surgery or change in appearance of tumour on follow up (n=18 [5.3%]). Only 8 tumours increased in size during follow up whilst the remaining 139 (94.6%) were the same size, smaller or completely resolved. (Table 3) In these 8 women the cyst increased in size between 21% and 94% but none of the women had surgery for that reason. In women

managed expectantly the median time for their follow-up ultrasound scan was 14.5 months (12 to 68 months maximum).

Using pattern recognition as the primary diagnostic test, 148 women were diagnosed with cancer. The management of women in respect of the predicted nature of pelvic tumour is shown in Table 4. The predicted diagnoses of the nature of the adnexal tumours at the initial visit using PR, LR1 and LR2 are shown in Table 5. There were significant differences in the proportion of women diagnosed with malignancy between PR and LR1/LR2. ($P<0.0001$) (Table 6) Assuming that the rate of intervention would be identical in women diagnosed with cancer and benign disease regardless of the diagnostic method used, there would also be a significant increase in the number of women having surgery for presumed malignancy if IOTA models were used instead of PR to assess women for interventions. ($P< 0.0001$) This is because the false positive rate for LR1/LR2 was significantly higher that of pattern recognition. The overall intervention rates including both benign and malignant lesions for PR, LR 1 and LR2, however would not be significantly different. ($P>0.05$) (Table 5)

There were 10 patients who had surgery after completing 12 months follow up which enabled a comparison of the two reference standards. In one of them there were discordant results for histology and the follow-up ultrasound scan. These results; however, suggest a good level of agreement between histology and the follow up visit strategy. The woman with the discordant results had

surgery because her tumour had changed morphologically. Histology result however was benign polypoid endometriosis.

The primary analysis was performed assuming that both reference tests (histology and follow up ultrasound scans) had the same perfect accuracy. (Table 7) This showed LR1, LR2 and pattern recognition all demonstrated high sensitivities, but there were significant differences in the specificities for the diagnosis of ovarian cancer between PR and the IOTA models when tests were used at the initial visit. When assuming differing accuracy for histology (100%) and follow up ultrasound scans (90%), the results were similar to when both reference tests were assumed to be 100% accurate. (Table 8)

There were eight women with cancer (borderline [5] and invasive [3]) who were misdiagnosed as benign disease using pattern recognition (8/137; 5.8% [95% CI: 2.6% to 11.2%]). Among the invasive malignancies, all were stage 1 with two dermoid tumours with early malignant transformation within and the one seromucinous adenocarcinoma. This caused delayed intervention in only one woman who was eventually diagnosed with a borderline tumour. Using LR1 there would have been four false negative cancer diagnoses (borderline [2] and invasive [2]) (2.9% [95%CI: 0.8% to 7.3%]) (P = 0.13 when compared to PR) diagnoses compared to seven (borderline [3] and invasive [4]) (5.1% [95%CI: 2.1% to 10.2%]) using LR2 (P=1.0 when compared to PR) neither of which was statistically significant. The tumours missed by LR1/LR2 were those missed by PR except for 1 metastatic tumour detected by PR and missed by LR1.

Discussion

Our study has shown that accuracy of LR1/LR2 for the diagnosis of ovarian cancer in the outpatient setting was similar to the previous studies, which were carried out pre-operatively. However, if LR1/LR2 had been used as the primary diagnostic test to guide the management decisions, rather than the pattern recognition, a significantly higher proportion of women would have been referred for treatment by gynaecological oncologists because of suspected ovarian cancer.

In previous original and subsequent validation studies the IOTA logistic regression models provided accurate diagnosis of ovarian cancer both in hands of expert and non-expert operators.[1,4,16-18] The models; however, had always been used in population of women who all had surgery. In view of that the results could be affected by selection bias and they cannot be extrapolated to low risk population majority of whom do not require surgical intervention. The results of previous studies can therefore only be used to help to select a surgeon (general or oncological) who should do the operation or the route of the surgery (laparoscopic or open). [19] A more relevant question in clinical practice; however, is whether surgery is required at all for the woman where symptoms, fertility or her wishes do not indicate surgery is required.

A difficulty in conducting studies on populations of women with adnexal tumours who are managed conservatively is the lack of agreement on the reference

standard to define the nature of the lesion. In women with presumed benign lesions the only way to rule out an ovarian cancer in the absence of histological diagnosis is by arranging follow up visits for a certain length of time. The natural history of ovarian cancer is unknown and the decisions and about the length and frequency of follow visits are pragmatic and based on consensus of opinions, rather than science. Most ovarian cancer screening projects adopted the policy of six-monthly or annual visits in women with normal ultrasound findings, which is deemed to be sufficiently frequent to detect early disease before spreading beyond the ovaries. [20] All women in our study had detectable lesions at the time of the initial scan. We therefore postulated that under these circumstance a 12 month follow up should be long enough to detect changes in the appearance and size of adnexal tumours which would be suggestive of their malignant nature. The absence of such changes was our second reference standard to discriminate between benign and malignant lesions.

During follow-up, it is possible that some women could develop new abnormalities. This may erroneously be classified as a prior misdiagnosis rather than the new disease that it is. However, in some women, it should be possible to identify two separate lesions which may overcome this limitation in a proportion of cases.

All the women in the study were assessed by at least one of the reference standards: histology and/or a follow-up ultrasound scan in12 months or more. This helped us to reduce the incomplete verification bias of including only

surgical patients. Bias may though still arise if the results of the alternative reference standard are treated identically to the results from the preferred reference standard when they may not be identical. This is because the two reference standards may be of different quality and it is possible that follow up ultrasound scans would be less than 100% accurate when compared to histology (preferred reference test). This information was taken into account when performing the statistical analysis, but in our study the results were not different when the presumed accuracy of ultrasound follow up (alternative test) was reduced to 90%. [14]

Our results showed that in this population with a 28.0% malignancy rate the LR1 and LR2 models had a high sensitivity and a moderate specificity to diagnose ovarian cancer. The specificity of LR1 and LR2 models were significant lower than pattern recognition ($P < 0.0001$) whilst the sensitivity was not significantly different ($P=0.13$ [LR1]). This relatively high false positive rate of LR1 would theoretically result in 62% more women being treated for potential ovarian cancer. However, the four women with false negative diagnoses of ovarian cancer (one invasive epithelial and three borderline) would have received earlier treatment.

Strengths –

This study assessed and corrected for the potential bias of analysing only the women who had surgery. This is the first study on LR1/LR2 to do this.

Limitations –

The overall intervention rate in our study was relatively high which reflects the nature of work in a large clinical centre. Many women are referred following the diagnosis of adnexal cyst in primary or community care. They often have larger lesions and are advised by their GPs that surgery is required even in the lesions are considered to be benign. The diagnostic strategy described in this study is appropriate for specialists making decisions about surgery. Further work is required to assess the suitability of the IOTA models to assist general practitioners and Level I operators in deciding who should be referred to specialist centres for further assessment.

In addition, 10% of women who were all diagnosed with benign tumours on pattern recognition did not complete follow up. There were no missed malignancies in 147 women with complete follow up data who were also managed expectantly. In view of that it is unlikely that the comparison of performance of IOTA models with pattern recognition would have been significantly different had all the women completed the study.

Further studies are required to assess the performance of IOTA models in hands of Level II sonographers receiving direct referrals from primary care and

reproductive health care physicians where the risk of ovarian cancer would be lower and the proportion of complex tumours difficult to classify on ultrasound would be less than in our population.

Acknowledgements

Sources of Financial Support

Dr Gareth Ambler received a proportion of funding from the Department of Health's NIHR Biomedical Research Centres funding scheme.

References

1. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L; International Ovarian Tumor Analysis Group; International Ovarian Tumor Analysis Group. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *J Clin Oncol* 2005; **23**: 8794-801.

2. EFSUMB. Minimum training recommendations for the practice of medical ultrasound. *Ultraschall Med* 2006; **27**: 79-105.

3. EFSUMB. Minimum training requirements for the practice of Medical Ultrasound in Europe. *Ultraschall in Med* 2010; **31**: 426–427.

4. 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. 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* 2010; **36**: 226-234.

5. Knottnerus JA, Dinant GJ. Medicine based evidence, a prerequisite for evidence based medicine. BMJ 1997; **315**: 1109-10.

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

7. Valentin L. Pattern recognition of pelvic masses by gray-scale ultrasound imaging: the contribution of Doppler ultrasound. Ultrasound Obstet Gynecol. 1999;**14**:338-47.

8. Yazbek J, Raju KS, Ben-Nagi J, Holland T, Hillaby K, Jurkovic D. Accuracy of ultrasound subjective 'pattern recognition' for the diagnosis of borderline ovarian tumors. *Ultrasound Obstet Gynecol* 2007; **29**: 489-495.

9. Sokalska A, Timmerman D, Testa AC, Van Holsbeke C, Lissoni AA, Leone FP, Jurkovic D, Valentin L. Diagnostic accuracy of transvaginal ultrasound examination for assigning a specific diagnosis to adnexal masses. *Ultrasound Obstet Gynecol* 2009; **34**: 462-70.

10. Yazbek J, Ameye L, Testa AC, Valentin L, Timmerman D, Holland TK, Van Holsbeke C, Jurkovic D. Confidence of expert ultrasound operators in making a diagnosis of adnexal tumor: effect on diagnostic accuracy and interobserver agreement. *Ultrasound Obstet Gynecol* 2010; **35**: 89-93.

11. Serov SF, Scully RE, Sobin LH. Histological Typing of Ovarian Tumours. (WHO International Histological Classification of Tumours No. 9). World Health Organization, Geneva, Switzerland. 1973.

12. Benedet JL, Bender H, Jones H 3rd,Ngan HY, Pecorelli S. FIGO staging classifications and clinical practice guidelines in the management of gynaecologic cancers. FIGO Committee on Gynecologic Oncology. *Int J Gynaecol Obstet* 2000; **70**: 209-262.

13. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, Springer, 2001.

14. de Groot JA, Dendukuri N, Janssen KJ, Reitsma JB, Bossuyt PM, Moons KG. Adjusting for differential-verification bias in diagnostic-accuracy studies: a Bayesian approach. *Epidemiology* 2011; **22**: 234-241.

15. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. *The BUGS Book - A Practical Introduction to Bayesian Analysis*, CRC Press, Chapman and Hall, London, 2012.

16. Nunes N, Ambler G, Hoo WL, Naftalin J, Foo X, Widschwendter M, Jurkovic D. A prospective validation of the IOTA logistic regression models (LR1 and LR2) in comparison to subjective pattern recognition for the diagnosis of ovarian cancer. *Int J Gynecol Cancer* 2013; **23**: 1583-1589.

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

18. Kaijser J, Sayasneh A, Van Hoorde K, Ghaem-Maghami S, Bourne T, Timmerman D, Van Calster B. Presurgical diagnosis of adnexal tumours using mathematical models and scoring systems: a systematic review and meta-analysis. *Hum Reprod Update* 2014; **20**: 449-462.

19. Campbell S. Ovarian cancer: role of ultrasound in preoperative diagnosis and population screening. *Ultrasound Obstet Gynecol* 2012; **40**: 245-254.

20. 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. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. 2016 Mar 5;**387**: 945-56.

Figure Legend

Figure 1 Flowchart of study participants and their management

Figure 1

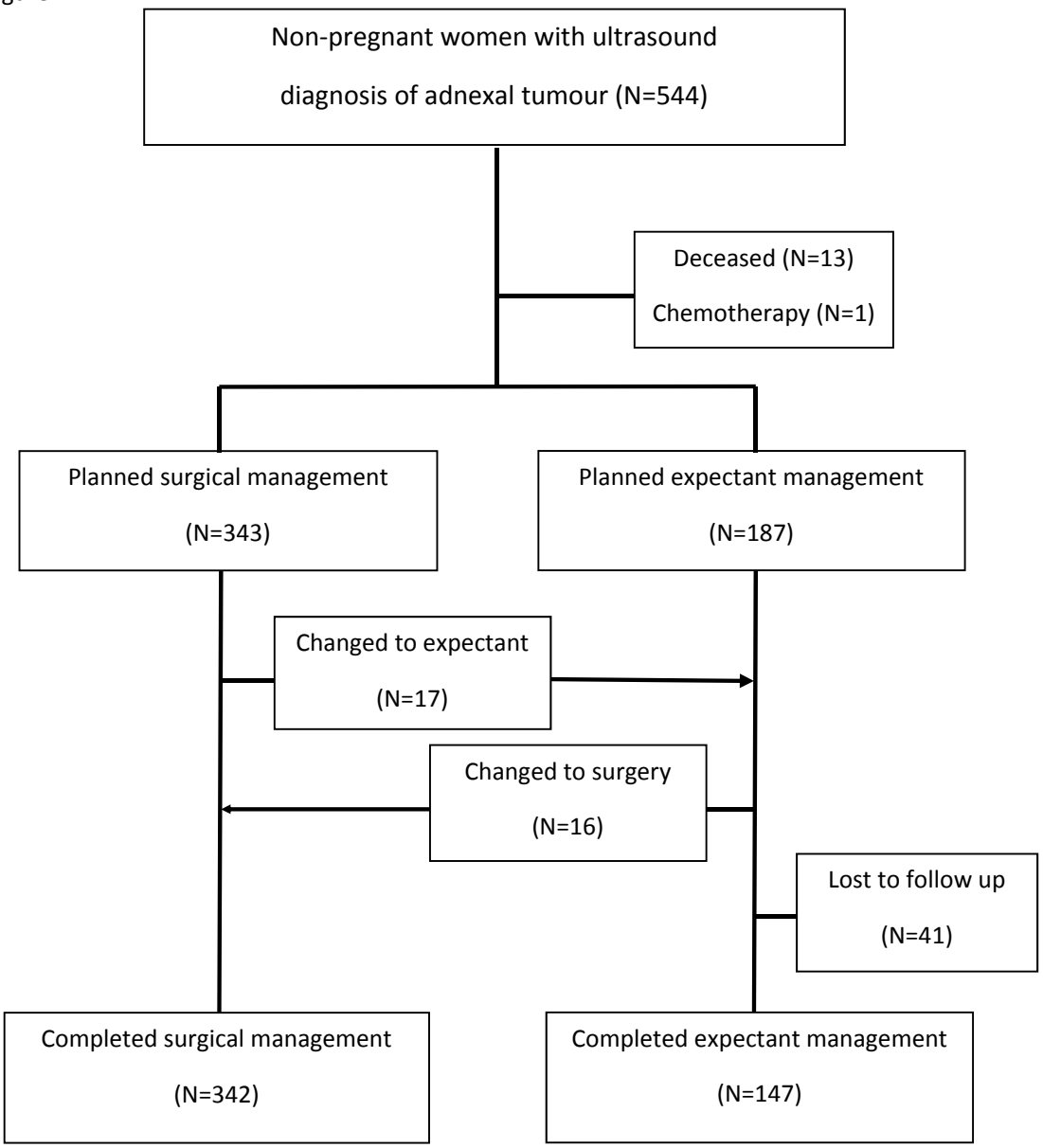

Figure 1 Flowchart of study participants and their management.

Table 1: Final Reference Standard Allocation in 544 non-pregnant women included into the study.

Reference Standard	N(%)
Histology Only	332 (61.0)
Follow-up USS Only	147 (27.0)
Both	10 (1.8)
Neither	55 (10.1)
Total	544 (100)

Table 2: Histological diagnoses (N=342).

Histology		N (%)
Benign n=205 (59.9%)	Cystadenoma (n=43) /Cystadenofibroma (n=14)	
	Endometrioma	57 (27.8)
	Mature cystic teratoma (Dermoid)	47 (22.9)
	Benign functional (haemorrhagic cyst, follicular cyst)	31 (15.1)
	Fibroma	29 (14.1)
	Normal adnexa at surgery	10 (4.9)
	Torsion (Functional x3, Cystadenofibroma, Fibroma)	5 (2.4)
	Pedunculated Leiomyoma	5 (2.4)
	Pelvic Inflammatory Disease	4 (2.0)
	Fibrothecoma	4 (2.0)
	Hydrosalpinx	3 (1.5)
	Pseudocyst	3 (1.5)
	Brenner	2 (1.0)
	Actinomycosis	2 (1.0)
	Struma Ovarii	1 (0.5)
	Myolipoma	1 (0.5)
Malignant n=137 (40.1%)	Invasive epithelial	84(61.3)
	Borderline (serous, mucinous, endometroid)	21(15.3)
	Invasive non-epithelial	4 (2.9)
	Metastases, Recurrence, Unknown	28(20.4)

Table 3: Morphological appearances of expectantly managed adnexal tumours on follow up scans (N = 147).

Follow up findings	N (%)
Resolved	34 (23.1)
Smaller*	21 (14.3)
Unchanged	84 (57.1)
Larger*	8 (5.4)
Total	147 (100)

*>20% change in the mean diameter

Table 4: Summary of management in respect to the predicted nature of adnexal tumour (N=544)

Management	Benign (N=396)	Malignant (N=148)
Surgery (N,%)	200 (50.5)	142 (95.9)
Chemotherapy (N,%)	-	1 (0.7)
Expectant (N,%)	147 (37.1)	-
Deceased (N,%)	9 (2.3)	4(2.7)
Lost to follow up (N,%)	40 (10.1)	1 (0.7)

Table 5: Proportion of women with ultrasound diagnosis of malignancy using pattern recognition, LR1 and LR2 and associated true and estimated*# intervention rates (N = 489).

	Malignant Diagnosis n	Intervention rates for Malignant Diagnoses n (%)	Intervention rates for All Diagnoses n (%)
Pattern Recognition	142	142/489 (29.0%)	342/489 (69.9%)
LR1	213	213/489 (43.6%)*	372/489 (76.1%)#
LR2	212	212/489 (43.4%)*	372/489 (76.1%)#

*Malignant diagnosis: Assumption that in keeping with pattern recognition all women with a malignant diagnosis using LR1 and LR2 would have had surgery.

#Benign Diagnosis: Assumption that in keeping with PR where 200/347 (57.6%) with benign diagnoses had surgery, the same proportion of women with benign diagnoses by LR1 and LR2 would also have had surgery.

#LR1 rate = (213 + 57.6% of 276)/489 = 76.1%

#LR2 rate = (212 + 57.6% of 277)/489 = 76.1%

Table 6: A comparison between LR1 and Pattern Recognition in classifying the tumours as benign or malignant (n = 489)

REF = Benign		**PR**	
		Benign	*Malign*
LR1	*Benign*	271	1
	Malign	68	12

P < 0.0001

REF = Malignant		**PR**	
		Benign	*Malign*
LR1	*Benign*	4	0
	Malign	4	129

P = 0.13 (Exact)

Table 7: Primary analysis using both reference standards (N = 489, malignancy rate 137/489 (28.0%)

	Sensitivity (%, 95% CI)	Specificity (%, 95% CI)	ROC area* (95% CI)	Diagnostic OR (95% CI)
PR	94.2 (88.8 to 97.4)	96.3 (93.8 to 98.0)	0.952 (0.930 to 0.974)	420 (172 to 1027)
LR1	97.1 (92.7 to 99.2)	77.3 (72.5 to 81.5)	0.872 (0.846 to 0.898)	113 (42.1 to 303)
LR2	94.9 (89.8 to 97.9)	76.7 (71.9 to 81.0)	0.858 (0.829 to 0.887)	61.1 (27.9 to 134)

* Based on single cutpoint

Both reference standards are assumed to have perfect performance of 100% accuracy.

Table 8: Secondary Bayesian analysis assuming 'high' diagnostic performance of follow up ultrasound (90% accuracy) (N = 489, malignancy rate 137/489 (28.0%)

	Sensitivity (%, 95% CrI*)	Specificity (%, 95% CrI)	ROC area** (95% CrI)	Diagnostic OR (95% CrI)
PR	93.8 (88.9 to 97.0)	96.4 (94.1 to 98.0)	0.948 (0.923 to 0.968)	430 (169 to 971)
LR1	96.7 (92.8 to 98.8)	77.5 (73.0 to 81.5)	0.868 (0.840 to 0.894)	116 (42.1 to 297)
LR2	94.5 (89.8 to 97.6)	76.8 (72.3 to 81.0)	0.854 (0.823 to 0.882)	61.1 (27.4 to 128)

*CrI = credible interval.
** Based on single cutpoint
Follow-up scan reference standard assumed to have 90% accuracy compared to histology reference standard assumed to have 100% accuracy.
Two Markov Chain Monte Carlo (MCMC) chains of 3000 samples were run in OpenBUGS with the first 1000 of each discarded as the burn-in period.