Interobserver Agreement in Describing the Ultrasound Appearance of Adnexal Masses and in Calculating the Risk of Malignancy Using Logistic Regression Models

Povilas Sladkevicius and Lil Valentin

Abstract

Purpose: To estimate interobserver agreement with regard to describing adnexal masses using the International Ovarian Tumor Analysis (IOTA) terminology and the risk of malignancy calculated using IOTA logistic regression models LR1 and LR2, and to elucidate what explained the largest interobserver differences in calculated risk of malignancy.

Experimental Design: One hundred and seventeen women with adnexal masses were examined with transvaginal gray scale and power Doppler ultrasound by two independent experienced sonologists who described the masses using IOTA terminology. The risk of malignancy was calculated using LR1 and LR2. A predetermined risk of malignancy cutoff of 10% indicated malignancy.

Results: There were 94 benign, four borderline, and 19 invasively malignant tumors. There was substantial variability between the two sonologists in measurement results and some variability in assessment of categorical variables (agreement 40%–98%, Kappa 0.30–0.91). Interobserver agreement when classifying tumors as benign or malignant was 84% (98/117), Kappa 0.68 for LR1, and for LR2 85% (99/117), Kappa 0.68. When using LR1 and LR2, the interobserver difference in calculated risk was ≥25 percentage units in 9% (11/117) and 12% (14/117) of tumors, respectively. Differences in assessment of wall irregularity, acoustic shadowing, color score, and color flow in papillary projections explained most of these largest differences.

Conclusions: Interobserver agreement in classifying tumors as benign or malignant using the risk of malignancy cutoff of 10% for LR1 and LR2 was good. However, because risk estimates may differ substantially between sonologists, one should be cautious with using the risk value for counseling patients about their individual risk. Clin Cancer Res; 21(3); 594–601. © 2014 AACR.

Introduction

One of the first successful attempts to use ultrasound to discriminate between benign and malignant adnexal masses was made by Granberg and colleagues (1). They classified adnexal masses into five categories: unilocular, unilocular solid, multilocular, multilocular solid, and solid tumors and found that unilocular cysts, unilocular solid cysts, and multilocular cysts were rarely malignant. Later, subjective interpretation of ultrasound images of adnexal masses—pattern recognition—proved to be an excellent method for discriminating between benign and malignant adnexal masses (2–5) and also for making a specific diagnosis (e.g., endometrioma, hydrosalpinx, and etcetera; refs. 3, 6, 7). As an alternative to pattern recognition, several research teams (8–10) created logistic regression models including clinical and ultrasound information to calculate the individual risk of malignancy in adnexal masses. Because of unclear definitions of many of the ultrasound variables included in these models, the International Ovarian Tumor Analysis (IOTA) group suggested standardized terms and definitions to be used when describing ultrasound images of adnexal masses (11). The IOTA group also created and validated several mathematical models in which these standardized terms and definitions were used to calculate the risk of malignancy for each individual adnexal mass (12–14). Of these models, the logistic regression models LR1 and LR2, including 12 and six variables, respectively (Table 1), were suggested to be suitable for use in clinical practice (13, 14). However, even when using standardized terms and definitions, ultrasound examiners may evaluate the features of an adnexal mass differently. There may also be variability in measurement results. This means that the risk of malignancy calculated by LR1 or LR2 may vary both within and between ultrasound examiners. We have shown that this is indeed the case when experienced ultrasound examiners analyze three-dimensional (3D) ultrasound volumes of adnexal masses (15). However, analysis of 3D ultrasound volumes does not necessarily reflect a situation where live examinations are performed.

The aims of this study were to estimate interobserver agreement when live ultrasound scans are performed with regard to (i) describing adnexal masses using the IOTA terminology, (ii) the risk of malignancy calculated using the IOTA logistic regression models LR1 and LR2, and (iii) to elucidate what explains large interobserver differences in calculated risk of malignancy.
Translational Relevance

The International Ovarian Tumor Analysis (IOTA) group has developed two logistic regression models (LR1 and LR2), including clinical and ultrasound variables for calculation of the risk of malignancy in adnexal masses. It has been suggested that LR1 and LR2 can be used to counsel patients about their individual risk of malignancy and so may have a role in personalized medicine. In this work, we found large interobserver differences (≥25 percentage units) in the calculated risk of malignancy in about 10% of cases. The differences were explained by ultrasound examiners interpreting ultrasound images differently. We suggest measures to improve interobserver agreement. Until better interobserver agreement in the calculated risk of malignancy using LR1 and LR2 has been shown, one should be cautious with using the risk estimate for individual patient counseling.

Table 1. Variables for which interobserver reproducibility was estimated

<table>
<thead>
<tr>
<th>Variables included in logistic regression models LR1 and LR2* or of importance for these models</th>
<th>mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum diameter of adnexal mass</td>
<td></td>
</tr>
<tr>
<td>Maximum diameter of largest solid component</td>
<td>mm</td>
</tr>
<tr>
<td>Maximum diameter of largest solid component</td>
<td>≤50, &gt;50 mm</td>
</tr>
<tr>
<td>Presence of an entirely solid adnexal mass</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Papillary projections present</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Irregular internal cyst walls</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Acoustic shadows</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Color Doppler signals in papillary projection</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Color score</td>
<td>1, 2, 3, 4</td>
</tr>
<tr>
<td>Tenderness of adnexal mass at scan</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Ascites</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Other IOTA variables</td>
<td></td>
</tr>
<tr>
<td>Continuous IOTA variables</td>
<td></td>
</tr>
<tr>
<td>Mean diameter of adnexal mass</td>
<td>mm</td>
</tr>
<tr>
<td>Mean diameter of largest solid component</td>
<td>mm</td>
</tr>
<tr>
<td>Height of the largest papillary projection</td>
<td>mm</td>
</tr>
<tr>
<td>Categorical IOTA variables</td>
<td></td>
</tr>
<tr>
<td>Type of tumor</td>
<td>Unilocular, unilocular solid, multilocular, multilocular solid, solid</td>
</tr>
<tr>
<td>Mean diameter of adnexal mass</td>
<td>≤40, 41-60, 61-80, 81-100, and &gt;100 mm</td>
</tr>
<tr>
<td>Number of cyst locules</td>
<td>0, 1, 2, 3, 4, 5, 6-10, 11-20, and &gt;20</td>
</tr>
<tr>
<td>Septum present</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Incomplete septum present</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Solid component</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Papillary projections</td>
<td>0, 1, 2, 3, ≥4</td>
</tr>
<tr>
<td>Echogenicity of cyst fluid</td>
<td>Anechoic, low level, ground glass, mixed, no cyst fluid</td>
</tr>
<tr>
<td>Color Doppler signals detectable</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Ovarian crescent sign</td>
<td>Yes, no</td>
</tr>
<tr>
<td>Diagnosis based on LR1 and LR2</td>
<td></td>
</tr>
<tr>
<td>Calculated risk of malignancy using LR1</td>
<td>%</td>
</tr>
<tr>
<td>Calculated risk of malignancy using LR2</td>
<td>%</td>
</tr>
</tbody>
</table>

Materials and Methods

The Ethics Committee of Lund University approved the study protocol. Informed consent was obtained from all participants or participant’s guardian, after the nature of the procedures had been fully explained.

This is a prospective observational study of real-time live ultrasound examinations of adnexal masses. Consecutive patients referred for an ultrasound examination and found to have an adnexal mass judged to need surgical removal were scanned according to the research protocol by sonologist 1 (P. Sladkevicius) as part of the clinical ultrasound examination. A second ultrasound examination was carried out before surgery by sonologist 2 (L. Valentin). Both examiners used the standardized IOTA examination and measurement technique and the IOTA terminology (11) to describe their ultrasound findings and noted their results in a dedicated paper form. Sonologist 2 was blinded to the results of sonologist 1. Information on the clinical variables included in LR1 and LR2 (personal history of ovarian cancer, current hormonal

Abbreviations: LR1, logistic regression model 1; LR2, logistic regression model 2.

*The risk of malignancy using the LR1 is derived as \( y = \frac{1}{1 + e^{-z}} \), where \( z = -6.7468 + 1.56985 (a) - 0.9983 (b) + 0.0326 (c) + 0.00841 (d) - 0.8577 (e) + 1.5513 (f) + 1.3737 (g) + 0.9281 (h) + 0.0496 (i) + 1.1421 (j) - 2.3550 (k) + 0.4916 (l), and e is the mathematical constant and base value of natural logarithms. The 12 variables included in LR1 are as follows: (a) personal history of ovarian cancer (yes = 1, no = 0), (b) current hormonal therapy (yes = 1, no = 0), (c) age of the patient (in years), (d) maximum diameter of the lesion (in millimeters), (e) mass tender at scan (yes = 1, no = 0), (f) the presence of ascites (yes = 1, no = 0), (g) the presence of blood flow within a solid papillary projection (yes = 1, no = 0), (h) the presence of a purely solid tumor (yes = 1, no = 0), (i) maximal diameter of the solid component (expressed in millimeters, but with no increase >50 mm), (j) irregular internal cyst walls (yes = 1, no = 0), (k) the presence of acoustic shadows (yes = 1, no = 0), and (l) color score (1, 2, 3, or 4). The risk of malignancy using LR2 is derived as \( y = \frac{1}{1 + e^{-z}} \), where \( z = -5.3718 + 0.0354 (c) + 1.6159 (f) + 1.1768 (g) + 0.0687 (i) + 0.9586 (l) - 2.9486 (k) \). A calculated risk of malignancy of more than 10% is classified as malignant (12).

1Includes 0.

1Includes only cases in which papillary projections were registered at both examinations.
therapy, age of the patient at menopause) was obtained at the preoperative ultrasound examination by sonologist 2. All patients were operated on within 90 days after the preoperative ultrasound examination performed by sonologist 2. The excised tissues underwent histologic examination and tumors were classified according to the criteria recommended by the International Federation of Gynecology and Obstetrics (16). Borderline tumors were classified as malignant.

The patients were examined in the lithotomy position with an empty urinary bladder (11). Abdominal ultrasound examination was added when needed. The ultrasound variables assessed with regard to interobserver reproducibility are shown in Table 1. The size of the lesion and that of its largest solid component were measured (largest diameter and mean of three orthogonal diameters) using calipers on the frozen ultrasound image. A color score was assigned on the basis of subjective assessment of the color content of the tumor scan at power Doppler ultrasound examination. A color score of 1 indicates absence of color Doppler signals, a color score of 2 a minimal amount of color Doppler signals, a color score of 3 a moderate amount of color Doppler signals, and a color score of 4 a large amount of color Doppler signals in the tumor (11).

The ultrasound systems used were GE Voluson 730 Expert or GE Voluson E8 (GE Healthcare) with a 5- to 9-MHz transvaginal transducer. For power Doppler ultrasound examinations, the following settings were used: for the Voluson 730 Expert system frequency, 6–9 ("normal") MHz; pulse repetition frequency, 0.6 kHz; gain 0.8; wall motion filter "low 1" (40 Hz); and for Voluson E8 frequency, 6–9 ("normal") MHz; pulse repetition frequency, 0.6 kHz; gain −4.0; wall motion filter "low 1" (40 Hz).

Statistical analysis

The IOTA3 study screen (astraia GMBH) was used to calculate the risk of malignancy according to LR1. Weighted Kappa indices were calculated using the statistical program Stata, Version 10.1 for Windows (StataCorp LP). For all other statistical calculations including calculation of the risk of malignancy when using LR2, we used the Statistical Package for the Social Sciences (SPSS program; IBM corp.; PASW version 18.0).

Interobserver agreement in the assessment of categorical variables was estimated by calculating the percentage agreement. Cohen's kappa was used to estimate by how much the observed agreement exceeded that expected by chance (17). Weighted kappa values are presented where appropriate (18). It has been suggested that Kappa values ≥0.81 indicate very good agreement beyond chance, kappa values between 0.61 and 0.80 good agreement beyond chance, kappa values between 0.41 and 0.60 moderate agreement beyond chance, kappa values between 0.21 and 0.40 fair agreement beyond chance, and kappa values ≤0.20 poor agreement beyond chance (19).

Interobserver reproducibility of measurement results, including the calculated risks of malignancy using LR1 and LR2, was described as the difference between two measurement results. The differences between the measured values were plotted against the mean of the two measurements (Bland–Altman plots) to assess the relationship between the differences and the magnitude of the measurements (20). Systematic bias between two measurements was estimated by calculating the 95% confidence interval (CI) of the mean difference (mean difference ± 2 SE). If zero lay within this interval, no bias was assumed to exist between the two measurements. Interobserver agreement was expressed as the mean difference and limits of agreement (20). Ninety-five percent of differences between any future measurements are estimated to fall between the lower and upper limit of agreement. Interobserver reliability of measurements results was estimated by calculating the intraclass correlation coefficient (ICC) using analysis of variance (two way random model—absolute agreement; this allows generalization of the results to a population of observers). The ICC indicates the proportion of the total variance in measurement results that can be explained by differences between the individuals examined. It depends both on the magnitude of measurement errors and the true heterogeneity in the population in which measurements are made. The more variable the population investigated, the greater the ICC and the less variable the population, the smaller the ICC (21). It has been suggested that ICC values ≥0.90 are needed for a test to be used in clinical practice (22).

The sensitivity and specificity with regard to malignancy of LR1 and LR2, calculated using the information of sonologists 1 and 2, were calculated.

Results

In all, 117 consecutive women with adnexal masses who underwent surgery were examined with ultrasound by the two sonologists as described above. Thirty-four women had bilateral adnexal masses. The most complex mass—or the largest one, if both masses had similar ultrasound morphology—was used in our statistical analysis, the mass to be included being selected retrospectively to ensure that both sonologists contributed the same mass (right or left) to the analysis. Thus, 117 adnexal masses from 117 patients constitute our study population. The women’s age ranged between 14 and 88 years (median 53), and 63 (54%) women were postmenopausal. There were 94 benign, four borderline, and 19 invasively malignant adnexal masses (Table 2).

The time elapsed between the ultrasound examination of sonologists 1 and 2 was median 61 days (10th and 90th percentiles 2 and 204) for the tumors with malignant histology and median 14 days (10th and 90th percentiles 2 and 31; range, 1–41) for the tumors with malignant histology. There was no relationship between the number of days between the scans and the differences in measurement results or interobserver agreement.
agreement for discrete variables (Supplementary Figs. S1–S5 and Supplementary Table S1).

Interobserver reproducibility of measurement results is shown in Table 3. Bland–Altman plots showed no clear trend for interobserver differences in measurement results to change with the magnitude of the measurement values. Limits of agreement were wide for all measurements. There was one systematic difference between the two sonologists, sonologist 1 who always performed the first examination) obtaining higher measurement values for the maximum diameter of the mass. The least reliable measurement was the height of the largest papillary projection.

Interobserver agreement when assessing categorical ultrasound variables is shown in Table 4. For most categorical ultrasound variables, interobserver agreement beyond chance was good or very good (19). Interobserver agreement beyond chance for variables included in LR1 or LR2 was poorest for color score (agreement 40%, weighted Kappa 0.36), presence of blood flow in papillary projection (agreement 90%, Kappa 0.48), irregular cyst wall (agreement 79%, Kappa 0.56), and acoustic shadowing (agreement 85%, Kappa 0.58).

Bland–Altman plots illustrating the relationship between the magnitudes of the estimated risk of malignancy calculated using LR1 and LR2 and the interobserver difference in calculated risk are shown in Fig. 1. The plots manifest a diamond shape, i.e., the interobserver differences are smallest for the lowest and highest risks, and they are very small for risks ≤2.5% and ≥95%. Logarithmic transformation of the data (20) did not substantially change the shape of the scatter plot. Therefore, we present our results as absolute interobserver differences in calculated risk (in percentage units), see Table 5. There were no systematic differences in calculated risks between the two sonologists, and reliability, reflected by the ICC values, was good (22), with ICC values for LR1 and LR2 being 0.911 and 0.832, respectively. When classifying tumors as having a risk of malignancy: when using LR1, the interobserver difference in calculated risk was <10.0 percentage units, and in four cases it was ≥25.0 percentage units. In the 18 cases where the two sonologists obtained different results with regard to malignancy when using LR2, the absolute interobserver difference in calculated risk ranged from 8.8 to 67.9 percentage units: in two of the 18 cases, the absolute interobserver difference in calculated risk was <1.0 percentage units, in ten cases it was 10.0 to 24.9 percentage units, and in six cases it was ≥25 percentage units.

The Bland–Altman plots (Fig. 1) illustrate that for some tumors there were substantial interobserver differences in the calculated risk of malignancy: when using LR1, the interobserver difference in calculated risk was ≥25 percentage units in 11 tumors (9% of all tumors), and when using LR2, the interobserver difference in calculated risk was ≥25 percentage units in 14 tumors (12% of all tumors). To elucidate which interobserver differences explained these largest interobserver differences in calculated risk, we scrutinized each case where the difference was ≥25 percentage units. The results are shown in Supplementary Tables S2 and S3. When using LR1, a discrepancy for one single categorical variable explained the difference in four of the 11 cases, while a discrepancy for two categorical variables explained the difference in one case (differences in measurements being <5 mm in these five cases). In six cases, there were differences in one or two categorical variables but also substantial differences (6–61 mm) in at least one measurement result. In no case was the large difference in calculated risk explained exclusively by differences in measurement results. The categorical variables judged differently by the two sonologists in these 11 cases were color score (n = 5), irregular cyst wall (n = 5), flow in papillary projection (n = 3), and acoustic shadowing (n = 2).

When using LR2, a discrepancy for one single categorical variable explained the large difference in calculated risk (≥25 percentage units) in eight of the 14 cases (differences in measurements being <5 mm in these eight cases), and in four of the eight cases, the sonologists judged acoustic shadowing differently. In five cases, there were differences in one categorical variable but also a substantial difference (9 mm–61 mm) in the measurement of the largest solid component. In yet another case there were differences in two categorical variables as well as in the measurement of the largest solid component. The categorical variables judged differently by the two sonologists in these 14 cases were acoustic shadowing (n = 5), irregular cyst wall (n = 5), ascites (n = 3), and flow in papillary projection (n = 2).

The sensitivity with regard to malignancy when using LR1 (10% risk cutoff) was 100% (23/23; 95% CI, 82–100%) for both...
sonologists, the specificity was 74% (70/94; 95% CI, 64–82) for sonologist 1 and 63% (59/94; 95% CI, 53–72) for sonologist 2. The sensitivity when using LR2 was 100% (23/23; 95% CI, 82–100) for sonologist 1 and 91% (21/23; 95% CI, 72–98) for sonologist 2, and the specificity was 75.5% (71/94; 95% CI, 65–84) for both sonologists.

Discussion

We have shown substantial interobserver variability in the results of measurements taken in adnexal masses (wide limits of agreement). Interobserver agreement beyond chance was very good or good for most categorical variables, but it was only moderate or fair for some. Interobserver agreement above chance was poorest for variables highly dependent on subjective evaluation and/or machine settings, i.e., color score, presence of color Doppler signals in papillary projections, irregular cyst walls, acoustic shadowing (all four variables being included in LR1 or LR2), echogenicity of cyst fluid, and ovarian crescent sign. Despite this there was good interobserver agreement when classifying tumors as benign or malignant using the predetermined risk of malignancy cutoff of 10%. However, in some cases there were substantial differences in the calculated risk of malignancy between the two sonologists, the difference being ≥25.0 percent-age units in 9% of all tumors when using LR1 and in 12% of all tumors when using LR2.

The strength of our study is that it provides new information. To the best of our knowledge, there is only one publication reporting on interobserver agreement with regard to describing ultrasound findings in adnexal masses using the IOTA terminology (11) when performing live ultrasound examinations (23). However, that study (23) evaluated interobserver reproducibility of the calculated risk of malignancy using LR1 or LR2 after live scanning.

It is a limitation of our study that up to 204 days elapsed between the scans of the two sonologists (up to 41 days for malignant masses). Because days elapsed between the scans, theoretically, the interobserver differences could be explained by the lesions having changed in size or morphology between the scans. We find this highly unlikely for the following reasons. First, there was no relationship between the differences in

### Table 4. Interobserver agreement for categorical variables included in the risk calculation models LR1 and LR2 and for other categorical variables used to describe adnexal masses

<table>
<thead>
<tr>
<th>Variables used in models LR1 and LR2</th>
<th>Agreement, %</th>
<th>Kappa value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence of entirely solid tumor (yes, no)</td>
<td>97 (114/117)</td>
<td>0.91</td>
</tr>
<tr>
<td>Maximum diameter of largest solid component (&lt;50 mm, &gt;50 mm)</td>
<td>95 (111/117)</td>
<td>0.83</td>
</tr>
<tr>
<td>Maximum diameter of largest solid component (&lt;50 mm, &gt;50 mm)</td>
<td>92 (56/61)</td>
<td>0.82</td>
</tr>
<tr>
<td>Tenderness of adnexal mass (yes, no)</td>
<td>97 (113/117)</td>
<td>0.78</td>
</tr>
<tr>
<td>Ascites (yes, no)</td>
<td>97 (113/117)</td>
<td>0.73</td>
</tr>
<tr>
<td>Papillary projections (yes, no)</td>
<td>87 (102/117)</td>
<td>0.65</td>
</tr>
<tr>
<td>Solid component (yes, no)</td>
<td>84 (98/117)</td>
<td>0.67</td>
</tr>
<tr>
<td>Irregular cyst wall (yes, no)</td>
<td>79 (92/117)</td>
<td>0.56</td>
</tr>
<tr>
<td>Acoustic shadows (yes, no)</td>
<td>85 (99/117)</td>
<td>0.58</td>
</tr>
<tr>
<td>Color Doppler signals in papillary projections (yes, no)</td>
<td>90 (105/117)</td>
<td>0.48</td>
</tr>
<tr>
<td>Color Doppler signals in papillary projections (yes, no)</td>
<td>86 (18/21)</td>
<td>0.71</td>
</tr>
<tr>
<td>Color score (1, 2, 3, 4)</td>
<td>40 (47/117)</td>
<td>0.36</td>
</tr>
<tr>
<td>Other variables used to describe adnexal masses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tumor type (unilocular, unilocular solid, multilocular, multilocular solid, solid)</td>
<td>97 (113/117)</td>
<td>0.70</td>
</tr>
<tr>
<td>Mean diameter of tumor (mm)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤40, 41–60, 61–80, 81–100, &gt;100</td>
<td>66 (77/117)</td>
<td>0.76</td>
</tr>
<tr>
<td>Number of locules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0, 1, 2, 3, 4, 5, 6–10, 11–20, &gt;20</td>
<td>58 (68/117)</td>
<td>0.79</td>
</tr>
<tr>
<td>&lt;10, &gt;10</td>
<td>94 (110/117)</td>
<td>0.80</td>
</tr>
<tr>
<td>Septum (yes, no)</td>
<td>88 (103/117)</td>
<td>0.76</td>
</tr>
<tr>
<td>Incomplete septum (yes, no)</td>
<td>98 (115/117)</td>
<td>1</td>
</tr>
<tr>
<td>Number of papillary projections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0, 1, 2, 3, ≥4</td>
<td>81 (95/117)</td>
<td>0.64</td>
</tr>
<tr>
<td>1, 2, 3, ≥4</td>
<td>67 (14/21)</td>
<td>0.63</td>
</tr>
<tr>
<td>4, &gt;4</td>
<td>96 (112/117)</td>
<td>0.68</td>
</tr>
<tr>
<td>Echogenicity of cyst fluid (anechoic, low level, ground glass, mixed, no fluid)</td>
<td>67 (78/117)</td>
<td>0.56</td>
</tr>
<tr>
<td>Color Doppler signals detectable (yes, no)</td>
<td>84 (98/117)</td>
<td>0.30</td>
</tr>
<tr>
<td>Ovarian crescent sign (yes, no)</td>
<td>78 (91/117)</td>
<td>0.49</td>
</tr>
<tr>
<td>Fluid in Douglas pouch (yes, no)</td>
<td>87 (102/117)</td>
<td>0.72</td>
</tr>
</tbody>
</table>

### Notes:
- aIncludes only cases in which solid components were registered at both examinations.
- bIncludes 0.
- cIncludes only cases in which papillary projections were registered at both examinations.
- dIncludes 1.
- eIncludes 0.
- fIncludes 2.
- gIncludes 3.
- hIncludes 4.
- iIncludes 5.
- jIncludes 6.
- kIncludes 7.
- lIncludes 8.
- mIncludes 9.
- nIncludes 10.
- oIncludes 11.
- pIncludes 12.
- qIncludes 13.
- rIncludes 14.
- sIncludes 15.
- tIncludes 16.
- uIncludes 17.
- vIncludes 18.
- wIncludes 19.
- xIncludes 20.
- yIncludes 21.
- zIncludes 22.
- AIncludes 23.
- BIncludes 24.
- CIncludes 25.
- DIncludes 26.
- EIncludes 27.
- FIncludes 28.
- GIncludes 29.
- HIncludes 30.
- IIncludes 31.
- JIncludes 32.
- KIncludes 33.
- LIncludes 34.
- MIncludes 35.
- NIncludes 36.
- OIncludes 37.
- PIncludes 38.
- QIncludes 39.
- RIncludes 40.
- SIncludes 41.
- TIncludes 42.
- UIncludes 43.
- VIncludes 44.
- WIncludes 45.
- XIncludes 46.
- YIncludes 47.
- ZIncludes 48.

Published OnlineFirst November 25, 2014; DOI: 10.1158/1078-0432.CCR-14-0906
measurement results and the number of days between the scans (Supplementary Figs. S1–S5). Nor was there a clear tendency for interobserver agreement for discrete variables to depend on the time between the scans (Supplementary Table S1). Second, one would expect a lesion and its components to increase in size with time, but sonologist 1 performing the first scan obtained higher measurement values than sonologist 2. Third, it is our experience after having performed gynecological scans for more than 20 years that the ultrasound morphology of both benign and malignant adnexal masses remains constant over time, that benign adnexal lesions grow slowly, and that malignant masses do not change appreciably in size even during 1 month of observation. Therefore, we believe that the discrepancies between the two sonologists reflect true interobserver differences and not a change of the masses over time. A second limitation is that we did not include estimation of the reproducibility of retrieving anamnestic information (current hormonal therapy, personal history of ovarian cancer), the anamnestic information collected by the second sonologist being used in all cases. It cannot be entirely excluded that patients would answer differently when asked by different sonologists, or that sonologists could interpret the answers of the patients differently.

Figure 1.
A, scatterplot showing the relationship between interobserver difference (observer 1 minus observer 2) in calculated risk and magnitude of calculated risk when using logistic regression model LR1. The plot manifests a diamond shape, the differences being smallest for the lowest and highest risks. For risks <2.5% and >95%, the differences are very small. LOA, limits of agreement. B, scatterplot showing the relationship between interobserver difference in calculated risk and magnitude of calculated risk when using logistic regression model LR2. The plot manifests a diamond shape, the differences being smallest for the lowest and highest risks.
differently. A third limitation is that we did not estimate intraobserver reproducibility. We considered four scans (two per sonologist) likely to be unacceptable to patients. For the same reason, only two sonologists were involved in this study, and our results are generalizable only to sonologists with a similar level of experience.

The results of this live scanning study are similar to those of another study in which the same sonologists assessed the same variables using 3D ultrasound volumes from adnexal masses in another tumor population (15). The similarity in results between the two studies is surprising, because the conditions when assessing 3D ultrasound volumes are different from those during a live scan. When evaluating ultrasound volumes, sonologists are exposed to the same ultrasound images, and so any interobserver difference should be explained exclusively by differences in interpreting the ultrasound information. During a live scan, there are more sources of bias. This could result either in poorer or better interobserver agreement than when 3D ultrasound volumes are assessed: poorer because ultrasound examiners are likely to use different machine settings and scanning conditions may change from one minute to another, better because the dynamic nature of live scanning facilitates discrimination between solid components and amorphous tissue.

Our results showed that two experienced sonologists agreed quite well in their classification of masses as benign or malignant using the 10% risk of malignancy cutoff of LR1 and LR2, and that the diagnostic performance of LR1 and LR2 with regard to discrimination between benign and malignant tumors was similar for the two sonologists and similar to that reported by others (14, 26–28). This is reassuring, because the main purpose of using model LR1 and LR2 is to classify tumors as benign or malignant. Potentially, however, LR1 and LR2 can be used not only to classify adnexal masses as benign or malignant but also to counsel a patient about her individual risk of malignancy (13). If to use the calculated risk for individual counseling, one must be reasonably certain not only that the estimated risk agrees well with the true risk (when externally validated both LR1 and LR2 underestimated the true risk especially in the risk interval 30%–70%; ref. 14), but also that the risk estimates are reproducible, i.e., that different examiners will obtain similar risk estimates. Our results show that risk estimates may differ substantially between experienced observers, the difference in estimated risk being ≥25.0 percentage units in 9% and 12% of cases when using LR1 and LR2, respectively. Interobserver agreement above chance was poorest for those variables in the models that are heavily dependent on subjective evaluation, i.e., color score, presence of color Doppler signals in papillary projections, irregular cyst walls, and acoustic shadowing. Indeed, differences in these explained most of the largest interobserver differences in calculated risk of malignancy. In models based on a few variables, changing values in only one variable may result in large differences in predicted risks, whereas a model with many variables is less vulnerable to a change in one or even few variables. Our results illustrate this (Supplementary Tables S2 and S3). When using LR2 (which includes six variables), a change in value for one single categorical variable explained an interobserver difference in calculated risk ≥25 percentage units in eight of 14 cases, whereas when using LR1 (which includes 12 variables), a change in value for one single categorical variable explained an interobserver difference in calculated risk ≥25 percentage units in only four of 11 cases. Acoustic shadowing is a strong variable in both LR1 and LR2 and has great impact on the calculated risk in LR2 with only six variables. In our hands as well as in those of Ruiz de Gauna and colleagues (23), interobserver agreement for acoustic shadowing was at most moderate. The interobserver agreement for color score was only fair in our study, and color score is an important variable in LR1.

To improve interobserver reproducibility of calculated risks based on LR1 and LR2, interobserver differences in descriptions and measurements of adnexal masses using the IOTA terminology and measurement technique need to be reduced. One way to achieve this could be by providing courses on and training in how to examine and describe adnexal masses using the IOTA terms. Interactive courses in which a large number of ultrasound images are discussed with the course participants are likely to be very valuable in this respect. More precise definitions of the IOTA terms, for example by providing ample imaging material, would probably also help improve interobserver agreement. Special attention should be given to the variables with poorest reproducibility, i.e., the color score, wall irregularity, acoustic shadowing, and detection of blood flow in papillary projections. Until better interobserver agreement in the calculated risk of malignancy using LR1 and LR2 has been shown, one should be cautious with using the risk estimates for individual patient counseling.

### Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

### Authors’ Contributions

Conception and design: P. Sladkevicius, L. Valentin

Development of methodology: P. Sladkevicius, L. Valentin

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): P. Sladkevicius, L. Valentin

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): P. Sladkevicius, L. Valentin

Writing, review, and/or revision of the manuscript: P. Sladkevicius, L. Valentin

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): P. Sladkevicius

Study supervision: L. Valentin

### Grant Support

This work was supported by the Swedish Medical Research Council (grant no. B0012201 and D0228201), funds administered by Skane University Hospital, Älmannas Juhdhusets i Malmö Stiftelse for bekämpande av cancer (the Malmö General Hospital Foundation for fighting against cancer), Landstingstämmandet regional forskning och ALF-model (i.e., two Swedish governmental grants from the region of Scania), and funds administered by Skane University Hospital. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received April 23, 2014; revised September 12, 2014; accepted October 6, 2014; published OnlineFirst November 25, 2014.
Agreement in Ultrasound Assessment of Ovarian Masses

References

Interobserver Agreement in Describing the Ultrasound Appearance of Adnexal Masses and in Calculating the Risk of Malignancy Using Logistic Regression Models

Povilas Sladkevicius and Lil Valentin


Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-14-0906

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2014/11/26/1078-0432.CCR-14-0906.DC1

Cited articles
This article cites 28 articles, 4 of which you can access for free at:
http://clincancerres.aacrjournals.org/content/21/3/594.full.html#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.