A Plea for AUC Confidence Intervals in Diagnosis Models used in Gynecology

Brahim Hamadicharef
Tiara #22-02, 1 Kim Seng Walk, Singapore 239403
Email: bhamadicharef@hotmail.com

***Abstract*—Over the last decade many studies in the gynecology literature have been investigating the performance of diagnosis models such as Univariate, Risk of Malignancy Index (RMI) and Logistic Regression (LR). Typical performance results are claimed in terms of sensitivity (SEN), specificity (SPE), accuracy (ACC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), with some studies als including Receiver Operating Characteristic (ROC) curve and its Area Under the Curve (AUC). It remains, however, that all these measures do not reflect any sample size and thus making it sometimes difficult to assess with confidence the true performance of these diagnosis models, in particular for small sample size. In this paper, we propose to use systematically, a ROC-based methodology that makes possible to calculate the Confidence Interval (CI) at each ROC point. The methodology is generic and robust to sample size, and based on Probability Density Function (PDF) without any assumption on the distribution. We illustrate its use on 6 recent studies and show that results with the additional AUC 95% CI contour is more adequate to compare the performance of these diagnosis models, especially with studies using different sample size.**

I. Introduction

Numerous studies, in the gynecology literature, provide performance of diagnosis models such as univariate models, Risk of Malignancy Index (RMI) [1][2] and Logistic Regression (LR) [3][4]. Performance results are typically presented as sensitivity (SEN), specificity (SPE), accuracy (ACC), Positive Predictive Value (PPV), Negative Predictive Value (NPV). Some studies include additional Receiver Operating Characteristic (ROC) analysis with a ROC curve, calculation of the Area Under the Curve (AUC) [5] and Standard Error (SE) [6]. It remains, however, that all these measure do not reflect any sample size and thus making it difficult to assess with confidence the true performance of the models and diagnosis. A methodology that would take into account sample size would also be very useful when comparing results from studies with different sample size.

In this paper we introduce a methodology originally developed by Tilbury [7] that allow to calculate, at a defined level of confidence (typically 0.05 for 95%), the Confidence Interval (CI) contour for each point of the ROC curve. We illustrate its use with recent studies taken from the gynecology literature presented by Yamamoto [8], Manjunath [9], Ma [10], Obeidat [2], Ulusoy [11], and Aslam [12].

The rest of the paper is organized as follow. In Section II we formulate the problem. The most commonly used diagnosis models in gynecology are presented in Section III. In Section IV we illustrate the novel approach with results from recent studies in gynecology. Finally, we conclude the paper in Section V.

TABLE I
Single threshold contingency

		Gold Standard	
		Malign	Benign
Diagnosis	Malign (Positive)	True Positive (TP)	False Positive (FP)
	Benign (Negative)	False Negative (FN)	True Negative (TN)
		nMal	nBen

II. Problem Formulation

Let us consider the following typical scenario. The outcome of a classifier for a 2-class medical diagnosis (benign or malignant, i.e. from the Ground Truth) can be of four types: True Positive (TP) when the tumor is malignant and diagnosed correctly, True Negative (TN) when the tumor is benign and diagnosed correctly, False Positive (FP) when the tumor is benign but diagnosed incorrectly as malignant, and False Negative (FN) when the tumor is malignant but diagnosed incorrectly as benign. These are shown in Table I. We know that the sample size of each groups is nBen=TN+FP and nMal=TP+FN. Using these counts, one can calculate he sensitivity and specificity performance measures, typically published in the literature together with sample size of the benign (nBen) and malignant (nMal) groups. As a reminder, sensitivity and specificity are defined as $\frac{TP}{TP+FN}$ and $\frac{TN}{TN+FP}$, respectively. As useful and informative these performance measure can be to the clinical gynecologist, they lack any sense/dimension of sample size. A sensitivity of 80% could be from different ratios with different sample sizes. Measures such as Positive Predictive Value (PPV) and Negative Predictive Value (NPV) are use in gynecology research and also suffer from the same sample issue issue. There is thus a need to provide confidence Interval (CI) (e.g. 95%CI), for each point of the ROC curve, that will allow for e.g. to compare results from studies with different sample size.

Tilbury [7] proposed a ROC-based methodology for performance evaluation of intelligent medical systems as a Bayesian approach to the AUC CI calculation, which was shown particularly suitable for small sample size (more details in [13]). WE

978-1-4244-6497-5/10/$26.00 ©2010 IEEE

TABLE II
SUMMARY OF RECENT STUDIES

Study	Sen / Spe	AUC	TP	TN	FP	FN
	(% / %)	AUC-L (95%CI)				
Yamamoto [8](nBen = 213 / nMal = 40)						
RMI1 (150)	85.0 / 83.1	0.840 (0.752)	34	177	36	6
RMI2 (200)	90.0 / 82.6	0.863 (0.781)	36	176	37	4
RMI3 (150)	85.0 / 83.1	0.840 (0.752)	34	177	36	6
RMI4 (450)	87.5 / 91.1	0.893 (0.809)	35	194	19	5
Manjunath [9](nBen = 55 / nMal = 93)						
RMI3 (200)	74.2 / 90.9	0.826 (0.743)	69	50	5	24
CA125 (150	73.1 / 92.7	0.829 (0.749)	68	51	4	25
UltraSound1	50.5 / 56.4	0.535 (0.432)	47	31	24	46
UltraSound2	43.0 / 89.1	0.661 (0.574)	40	49	6	53
Post-Menopause	48.4 / 65.5	0.569 (0.467)	45	36	19	48
Ma [10](nBen = 77 / nMal = 63)						
RMI (400)	79.4 / 94.8	0.871 (0.792)	50	73	4	13
CA125 (200)	65.1 / 94.8	0.799 (0.715)	41	73	4	22
UltraSound	93.7 / 83.1	0.884 (0.809)	59	64	13	4
Post-Menopause	55.6 / 79.2	0.674 (0.577)	35	61	16	28
Obeidat [2](nBen = 28 / nMal = 72)						
RMI (200)	90.3 / 89.3	0.898 (0.793)	65	25	3	7
CA125 (300)	84.7 / 71.4	0.781 (0.656)	61	20	8	11
UltraSound	80.6 / 57.1	0.688 (0.561)	58	16	12	14
Post-Menopause	77.8 / 53.6	0.657 (0.529)	56	15	13	16
Ulusoy [11](nBen = 190 / nMal = 106)						
Age (54)	35.8 / 82.1	0.590 (0.525)	38	156	34	68
RMI (153)	76.4 / 77.9	0.772 (0.706)	81	148	42	25
CA125 (80)	65.1 / 82.1	0.736 (0.668)	69	156	34	37
UltraSound1	100.0 / 77.4	0.887 (—–)	106	147	43	0
UltraSound2	8.5 / 66.8	0.377 (0.325)	9	127	63	97
Post-Menopause	46.2 / 67.9	0.571 (0.499)	49	129	61	57
Aslam [12](nBen = 67 / nMal = 33)						
LR1 [3]	45.5 / 92.5	0.690 (0.581)	15	62	5	18
LR2 [15]	9.1 / 98.5	0.538 (0.484)	3	66	1	30
LR3 [4]	72.7 / 91.0	0.819 (0.706)	24	61	6	9
LR4=LR1+LR2	60.6 / 92.5	0.766 (0.652)	20	62	5	13
LR5=LR1+LR3	78.8 / 94.0	0.864 (0.757)	26	63	4	7
LR6=LR2+LR3	72.7 / 83.6	0.782 (0.663)	24	56	11	9
LR7=All	78.8 / 94.0	0.864 (0.757)	26	63	4	7

use a fast computation formulation for calculating the lower and upper bounds of AUC CIs, recently proposed in [14].

III. DIAGNOSIS MODELS

The most common diagnosis models in ovarian cancer research include univariate, Risk of Malignancy Index (RMI) and Logistic Regression (LR). Univariate models take a single variable such as post-menopause score, ultrasound score, serum CA125 level as unique feature and a diagnosis decision is taken by a threshold. Risk of Malignancy Index (RMI), first introduced by Jacobs [16], combines variables as follow:

$$RMI = U \times M \times CA125 \tag{1}$$

where M post-menopause score (M = 1 when patient is pre-menopause and M = 3 when patient is post-menopause), U is the ultrasound score (typically a total ultrasound score of 0 or 1 yielded U = 1, and a score of ≥2 yielded U = 4 [17]) and CA125 is a direct measure, typically in ml, of the level of CA125 serum. The RMI model, simple in its calculation and interpretation, has been the subject of numerous studies [18][17][19][1].

In recent years, Logistic Regression (LR) models have gained popularity in gynecology [20][21]. It aims to predict an outcome from multiple variables, using a form as:

$$y(X) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{i=1}^{N} \beta_i x_i\right)}} \tag{2}$$

where N the number of variables, X is set of variables, and β_0 ... β_N are the regression coefficients estimated using maximum-likelihood and the least-squares regression fitting procedures.

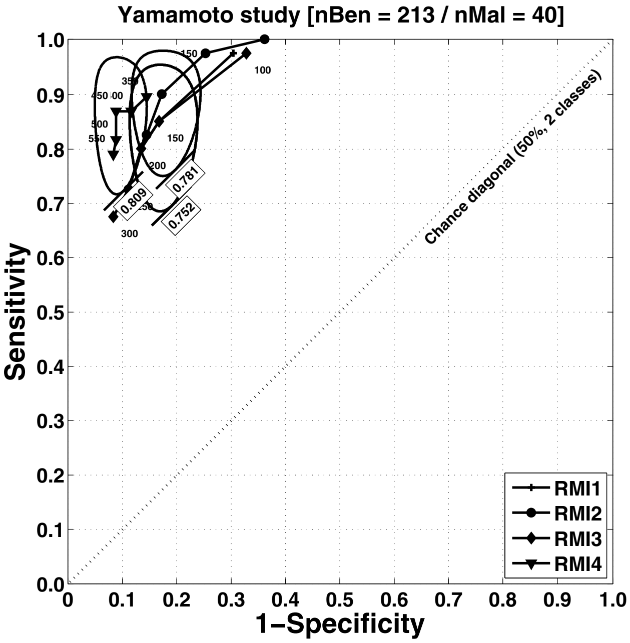

Fig. 1. Results of Yamamoto's study [8](N=253)

IV. RESULTS FROM RECENT STUDIES

We re-examined the results from few recent studies using our approach. For each model of each study, the sensitivity and specificity are summarized in Table II. Using our approach we calculate AUC at the ROC point and AUC-L, the 95% CI AUC lower bound. We also indicate the sample size in each group, with number of benign (nBen) and malignant (nMal) cases.

ROC curves for each models in each study are used to assess the 95% CI contour. ROC curve with 95% CI contours

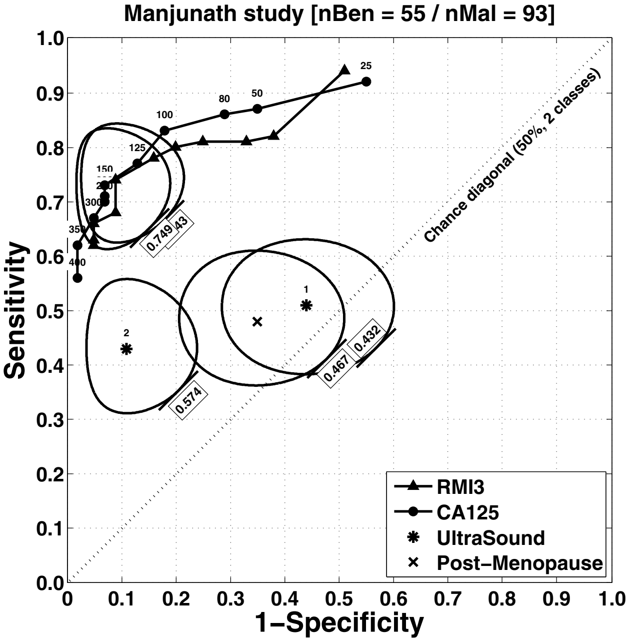

Fig. 2. Results of Manjunath's study [9](N=148)

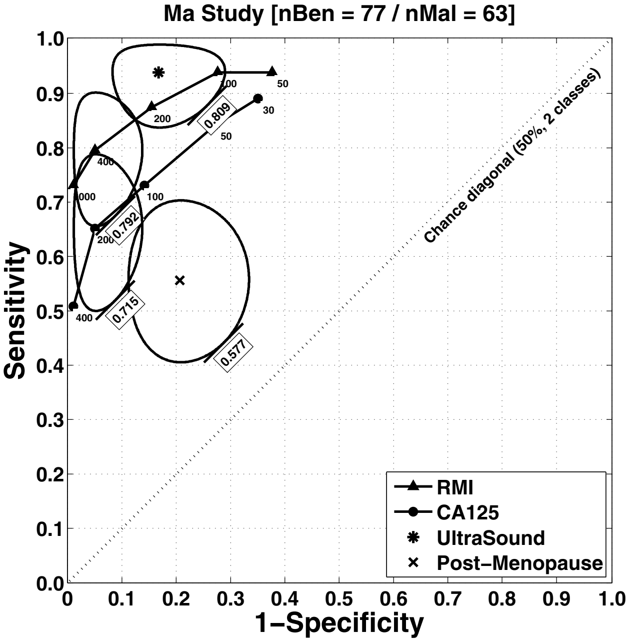

Fig. 3. Results of Ma's study [10](N=140)

are plotted for each model. Results from the Yamamoto's study [8] are presented in Figure 1, results from the Manjunath's study [9] are shown in Figure 2, from Ma's study [10] in Figure 3, those of Obeidat's study [2] in Figure 4, of Ulusoy's study [11] in Figure 5, and finally, results from the Aslam's study [12] are shown in Figure 6.

Due to the different performance and differences in sample size (both in terms of total number but also different ratio between benign and malignant), the 95% CI contours are large

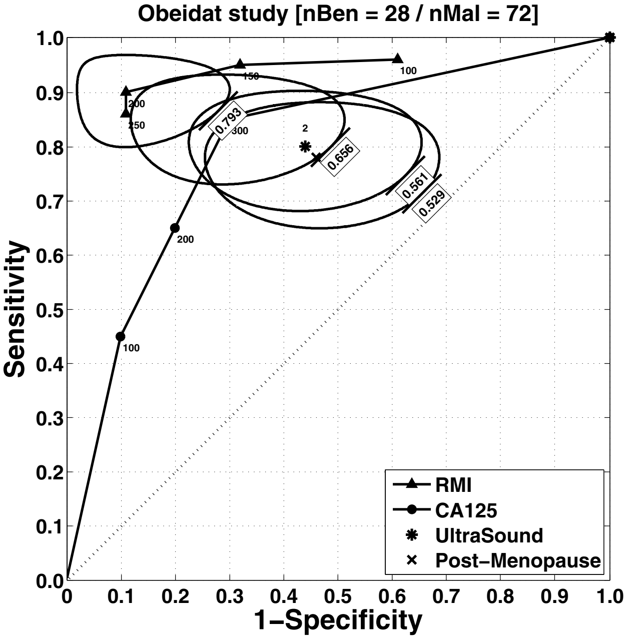

Fig. 4. Results of Obeidat's study [2](N=100)

and stretched. Some of the model have the contour close or even overlapping the level of chance (diagonal line) and thus their performance should be considered with more care.

V. Conclusions

In this paper, we presented a novel approach to visualize the performance of medical diagnosis models that overcome the issue of different sample size. It is recognized that as performance evaluation is typical claimed in terms of the sensitivity and specificity, and that, however, the sample size is rarely taken into consideration. Using the CI contour at the ROC, the methodology provide adequate visual indication as true performance (using the lower bound of the CI contour). We illustrated the methodology with the re-examination of the results from recently published studies which compared the performance of univariate, RMI and LR models for the diagnosis of ovarian cancer. Using the ROC with CI contour, one can assess the true performance of these diagnosis models, and furthermore, can assess any overlap of the CI contour with chance line and compare models all together considering that the difference in sample size is taken into account by the CI contour calculation.

Finally, it should be reminded that only the use of larger data set, such as the one created for the International Ovarian Tumor Analysis (IOTA) group with 1066 (800 benign and 266 malignant) patients [22], will give make the CI contour shrink and provide good confidence in the diagnosis performance.

References

[1] S. Ma, K. Shen, and J. Lang, "A risk of malignancy index in preoperative diagnosis of ovarian cancer," *Chinese Medical Journal*, vol. 116, no. 3, pp. 396–399, 2003.

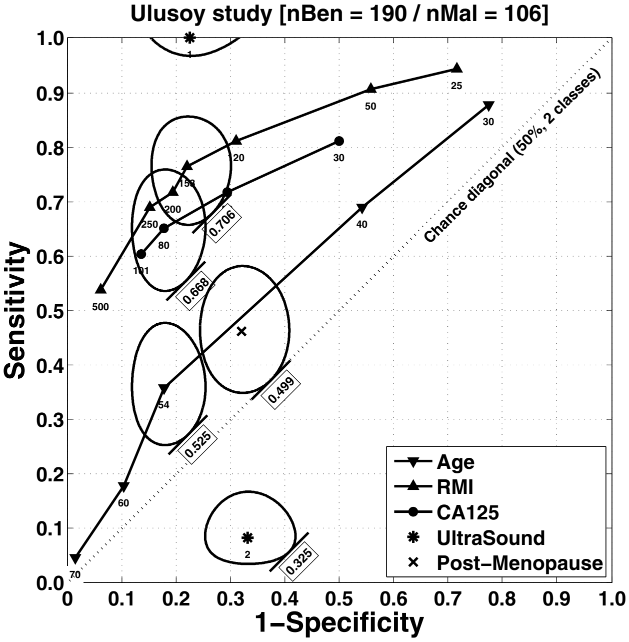

Fig. 5. Results of Ulusoy's study [11](N=296)

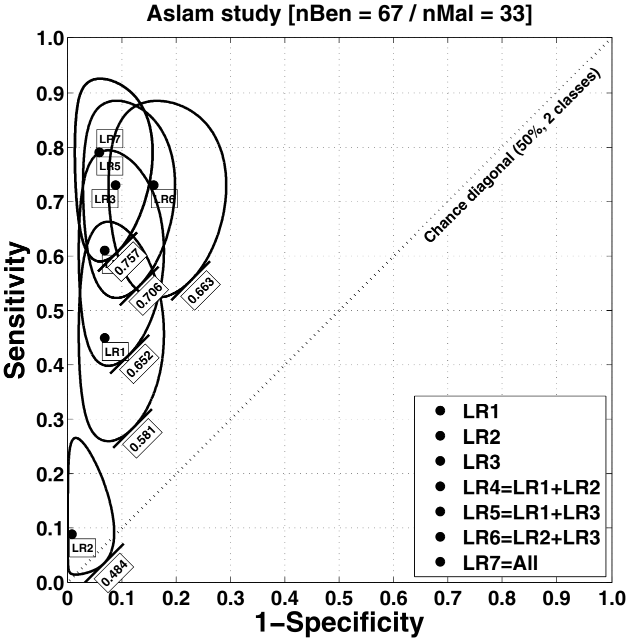

Fig. 6. Results of Aslam's study [12](N=140)

[2] B. R. Obeidat, Z. O. Amarin, J. A. Latimer, and R. A. Crawford, "Risk of malignancy index in the preoperative evaluation of pelvic masses," *International Journal of Gynecology & Obstetrics*, vol. 85, no. 3, pp. 255–258, June 2004.

[3] A. Tailor, D. Jurkovic, T. H. Bourne, W. P. Collins, and S. Campbell, "Sonographic prediction of malignancy in adnexal masses using multivariate logistic regression analysis," *Ultrasound in Obstetrics and Gynecology*, vol. 10, no. 1, pp. 41–47, July 1997.

[4] D. Timmerman, H. Verrelst, T. H. Bourne, B. De Moor, W. P. Collins, I. Vergote, and J. Vandewalle, "Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses," *Ultrasound in Obstetrics and Gynecology*, vol. 13, no. 1, pp. 17–25, January 1999.

[5] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, no. 1, pp. 29–36, April 1982.

[6] ——, "A Method of Comparing the Areas under ROC curves derived from same cases," *Radiology*, vol. 148, pp. 839–843, 1983.

[7] J. Tilbury, P. Van-Eetvelt, J. Garibaldi, J. Curnow, and E. Ifeachor, "Receiver Operator Characteristic Analysis for Intelligent Medical Systems - A New Approach for Finding Confidence Intervals," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 952–963, July 2000.

[8] Y. Yamamoto, R. Yamada, H. Oguri, N. Maeda, and T. Fukaya, "Comparison of four malignancy risk indices in the preoperative evaluation of patients with pelvic masses," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 144, pp. 163–167, 2009.

[9] A. P. Manjunath, Pratapkumar, K. Sujatha, and R. Vani, "Comparison of Three Risk of Malignancy Indices in Evaluation of Pelvic Masses," *Gynecologic Oncology*, vol. 81, no. 2, pp. 225–229, May 2001.

[10] S. Ma, K. Shen, and J. Lang, "A risk of malignancy index in preoperative diagnosis of ovarian cancer," *Chin Med J*, vol. 116, no. (Suppl 3), pp. 396–399, 2003.

[11] S. Ulusoy, O. Akbayir, C. Numanoglu, N. Ulusoy, E. Odabas, and A. Gulkilik, "The risk of malignancy index in discrimination of adnexal masses," *International Journal of Gynecology and Obstetrics*, vol. 96, no. 3, pp. 186–191, March 2007.

[12] N. Aslam, S. Banerjee, J. Carr, M. Savvas, R. Hooper, and D. Jurkovic, "Prospective Evaluation of Logistic Regression Models for the Diagnosis of Ovarian Cancer," *Obstetrics & Gynecology*, vol. 96, no. 1, pp. 75–80, July 2000.

[13] J. B. Tilbury, "Evaluation of Intelligent Medical Systems," Ph.D. thesis, Department of Communications and Electronic Engineering (DCEE), University of Plymouth, Drake Circus, Plymouth PL4 8AA, Devon, United Kingdom, September 2002.

[14] B. Hamadicharef, "Frequentist versus Bayesian approaches for AUC Confidence Interval Bounds," *Proceedings of the 10th International Conference on Information Science, Signal Processing and their applications (ISSPA2010), Kuala Lumpur, Malaysia, May 10–13, 2010*, pp. 341–344.

[15] J. L. Alcazar and M. Jurado, "Using a logistic model to predict malignancy of adnexal masses based on menopausal status, ultrasound morphology, and color Doppler findings," *Gynecology Oncology*, vol. 69, pp. 146–150, 1998.

[16] I. Jacobs and *et al*, "A risk of malignacy index incorporating Ca125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer," *British Journal of Obstetrics and Gynaecology*, vol. 97, pp. 922–929, 1990.

[17] S. Tingulstad, B. Hagen, F. E. Skjeldestad, M. Onsrud, T. Kiserud, T. Halvorsen, and K. Nustad, "Evaluation of a risk of malignancy index based on serum Ca125, ultrasound findings and menopausal status in the pre-operative diagnosis of pelvic masses," *British Journal of Obstetrics and Gynaecology*, vol. 103, no. 8, pp. 826–831, August 1996.

[18] J. Bailey, A. Tailor, R. Naik, A. Lopes, K. Godfrey, H. M. Hatem, and J. Monaghan, "Risk of malignancy index for referral of ovarian cancer cases to a tertiary center: does it identify the correct cases?" *International Journal of Gynecological Cancer*, vol. 16, pp. 30–34, February 2006.

[19] I. Jacobs, S. Skates, A. P. Davies, R. Woolas, A. Jeyerajah, P. Weidemann, K. Sibley, and D. Oram, "Risk of diagnosis of ovarian cancer after raised serum CA 125 concentration: a prospective cohort study," *British Medical Journal*, vol. 313, no. 7069, pp. 1355–1358, November 1996.

[20] K. S. Khan, P. F. W. Chien, and L. S. Dwarakanath, "Logistic Regression Models in Obstetrics and Gynecology Literature," *Obstetrics and Gynecology*, vol. 93, no. 6, pp. 1014–1020, June 1999.

[21] D. Timmerman, T. Bourne, A. Tailor, W. P. Collins, H. Verrelst, K. Vandenberghe, and I. Vergote, "A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: the development of a new logistic regression model," *American Journal of Obstetrics and Gynecology*, vol. 181, no. 1, pp. 57–65, July 1999.

[22] L. Valentin, L. Ameye, A. Testa, F. Lcuru, J.-P. Bernard, D. Paladini, S. Van Huffel, and D. Timmerman, "Ultrasound characteristics of different types of adnexal malignancies Gynecologic Oncology," *Gynecologic Oncology*, vol. 102, no. 1, pp. 41–48, July 2006.