Ultrasound Obstet. Gynecol. 10 (1997) 192–197

Transvaginal ultrasonographic characterization of ovarian masses: comparison of five scoring systems in a multicenter study

E. Ferrazzi, G. Zanetta, D. Dordoni†, N. Berlanda, R. Mezzopane and G. Lissoni**

Departments of Obstetrics and Gynecology, ISBM San Paolo and *ISBM San Gerardo, University of Milan; †University of Brescia, Italy

Key words: ULTRASOUND, TRANSVAGINAL ULTRASONOGRAPHY, ADNEXAL MASSES, OVARIAN CANCER, MORPHOLOGICAL SCORES

ABSTRACT

The aim of this work was to test and compare the accuracy of five different morphological scoring systems to identify malignant ovarian masses in a prospective multicenter study. Four of the systems had previously been reported by Granberg, Sassone, De Priest and Lerner and the fifth is newly developed.

A total of 330 ovarian neoplasms were collected in three different centers, which adopted the same diagnostic procedures. Of these, 261 masses were benign (mean diameter 50 ± 26 mm) and 69 were malignant (mean diameter 69 ± 33 mm) (prevalence 21%). The area under the receiver operating characteristic (ROC) curve for the multicenter score was 0.84. This was significantly better than the areas of the other four scores which ranged from 0.72 to 0.75. The cut-off levels derived from the five ROC curves achieved a sensitivity that ranged from 74% (Sassone score) to 88% (De Priest score ≥ 5), and a specificity from 40% (De Priest) to 67% (multicenter); the highest positive predictive value was 41% (multicenter). With a cut-off level of 9, the accuracy of the multicenter score was significantly better than the scores of Granberg and De Priest (McNemar's test $p < 0.0001$*). Similar results were obtained in 207 ovarian masses of ≤ 5 cm in mean diameter, and when 19 borderline and 11 stage I cancers only were considered. For the clinical purposes of a screening test we also checked a possible cut-off level of ≥ 8, which increased the sensitivity to 93% with a drop of specificity to 56%. With the use of the same criteria for the scores of the different authors, the following values were obtained for sensitivity: 96%, 81%, 93% and 90%; and for specificity: 23%, 56%, 28% and 49%.*

The multicenter score performed well at distinguishing malignant from benign lesions, and was better than the other four traditional scores, for both large and small masses. This was mainly due to the introduction of two criteria that allowed correction for typical dermoids and endohemorrhagic corpora lutea. A completely reliable differentiation of benign from malignant masses cannot be obtained by sonographic imaging alone.

INTRODUCTION

The introduction of transvaginal sonography has greatly changed the negative attitude of many gynecological oncologists[1] towards the possible role of ultrasound in the diagnosis of pelvic adnexal masses[2–4]. In many cases, the high resolution of this imaging technique apparently enables us to define the gross anatomy of ovarian lesions. However, the extreme variability in the macroscopic characteristics of benign and malignant lesions observed in large series prevents the achievement of precise pathological diagnoses from detailed sonographic descriptions.

To overcome these limitations of reproducibility, considerable effort has been made to identify a morphological scoring system[5–8], or morphological classes[2], which could find the most efficient cut-off level or allocation to categorize tumors as benign or malignant. The introduction of color Doppler, with its potential to identify neoangiogenesis in neoplasms[9–13], to a large extent has diverted the attention of researchers away from the advantages of a simple and reproducible morphological scoring system. In spite of the achievements of color and power Doppler imaging, morphological assessment is still used as the basis of ultrasound evaluation of ovarian lesions[8,14–16]. However, to our knowledge, the performance of these scores has never been compared. Furthermore, most tests were introduced with the assumption that they were an improvement on previous tests, without this being checked.

Correspondence: Dr E. Ferrazzi, San Paolo Biomedical Sciences Institute, Via A. di Rudini 8, 20142 Milan, Italy

Received 28–10–96
Revised 12–5–97
Accepted 28–5–97

The objective of this work was to compare the performance of four different scoring systems[2,5–7] prospectively in a series collected in three different centers and to test and compare the results of a new scoring system developed by these centers, with the aim of overcoming the limitations of the previously used scores.

SUBJECTS AND METHODS

This study was performed in three Departments of Gynecology of the Universities of Milan (ISBM San Paolo, center 1; and ISBM San Gerardo, center 2) and Brescia (center 3). Over a 2-year period in 1995 and 1996, 317 consecutive patients were recruited. A total of 330 adnexal masses were included in this study: 138 examined in center 1, 119 in center 2 and 73 in center 3. The mean age (± SD) in this series was 45 years (± 16), and the range was from 19 to 89 years. Of these patients, 112 were postmenopausal and 205 were premenopausal. The modal age for benign lesions was 40 years (interquartile range 29–48 years), and the modal age for malignant lesions was 49 years (interquartile range 35–62 years). Inclusion criteria were surgery within 7 days after ultrasound diagnosis, and a detailed pathological report of the lesion.

Five different ultrasound machines were used, and each one had multifrequency (5–7-MHz) transvaginal transducer probes. Nine gynecologists with experience in gynecological ultrasound varying from 1 year to more than 5 years performed the scans. The mean diameter of each mass was calculated according to a simple geometric formula (longitudinal + anteroposterior + transverse diameters divided by 3). For every lesion a comprehensive set of images was recorded.

Five different morphological scoring systems were used to quantify the risk of malignancy. Four of these scores were chosen from those previously published, by Granberg and colleagues[2], Sassone and colleagues[5], De Priest and colleagues[6] and Lerner and colleagues[7] (Table 1). The fifth scoring system (Table 2), developed by our centers, was based on the following criteria: (1) a guideline which could be used by the sonologist/sonographer to provide a systematic description of the lesion; (2) a non-continuous point value assignment which could take into account the different pathological correlates of macroscopic aspects of neoplasms with regard to their potential malignancy, as suggested by the work of Lerner and co-workers[7]; (3) fine trabecular jelly-like content, typical of endohemorrhagic corpus luteum given the same low value as anechogenic content; and (4) a diagnostic system which could easily solve the problem encountered by the high scores usually achieved by benign cystic teratomas frequently met in young women and which in no way can be considered as being at high risk of malignancy. When a typical cystic teratoma was observed (heterogeneous mass with

Table 1 Detailed description of the four previously published scoring systems

	Score				
Granberg et al.[2]					
Morphology	0	1	2	3	4
Ultrasound finding	unilocular	unilocular solid	multilocular	multilocular solid	solid
Sassone et al.[5]					
Morphology	1	2	3	4	5
Inner wall structure	smooth	irregularities ≤ 3 mm	papillarities > 3 mm	not applicable, mostly solid	—
Wall thickness (mm)	thin (≤ 3)	thick (> 3)	not applicable, mostly solid	—	—
Septa (mm)	none	thin (≤ 3)	thick (> 3)	—	—
Echogenicity	sonolucent	low echogenicity	low echogenicity with echogenic core; mixed echogenicity		high echogenicity
De Priest et al.[6]					
Morphology	0	1	2	3	4
Cystic wall structure	smooth (< 3 mm thick)	smooth (> 3 mm thick)	papillary projection (< 3 mm)	papillary projection (≥ 3 mm)	predominantly solid
Volume (cm^3)	< 10	10–50	> 50–200	> 200–500	> 500
Septum structure	no septa	thin septa (< 3 mm)	thick septa (3 mm to 1 cm)	solid area (≥ 1 cm)	predominantly solid
Lerner et al.[7]					
Morphology	0	1	2	3	
Wall structure	smooth or small irregularities < 3 mm	—	solid or not applicable	papillarities ≥ 3 mm	
Shadowing	yes	no	—	—	
Septa	none or thin (< 3 mm)	thick (≥ 3 mm)	—	—	
Echogenicity	sonolucent or low-level echo or echogenic core	—	—	mixed or high	

Table 2 'Multicenter' scoring system

Score	*Wall*	*Septa*	*Vegetations*	*Echogenicity*
1	≤ 3 mm	none	none	sonolucent*
2	> 3 mm	≤ 3 mm		low echogenicity
3		> 3 mm		
4	irregular, mostly solid†		≤ 3 mm	with echogenic areas
5	irregular, not applicable‡		> 3 mm	with heterogeneous echogenic areas, solid

*Or with fine trabecular and jelly-like hypoechoic content typical of endohemorrhagic corpus luteum; †irregular wall structure, much thicker than 3 mm but capsule identifiable; ‡the capsule cannot be differentiated from the surrounding structure and the inner echogenicity

irregular hypoechoic and hyperechoic areas with posterior shadowing, not separated by septa, or homogeneous hyperechoic mass with regular capsule and posterior shadowing), the scoring system was disregarded, and the lesion was given the lowest possible value.

The five scores adopted were compared by calculation of the receiver operating characteristic (ROC) curve for each of them in the whole series. In a similar way, ROC curves were calculated for lesions with a mean diameter of ≤ 5 cm. The ROC curves obtained for the whole series by the scores of Granberg, Sassone and De Priest and their co-workers were also calculated, with the correction described for typical cystic teratomas; the 'corrected' values which were assigned to these systems were 0, 4 and 0, respectively. The area under each ROC curve was measured, as described by Hanley and McNeill[17]. When each score was compared with the other four in the series, we obtained ten comparisons of paired samples. (Score 1 was compared with scores 2, 3, 4 and 5 (four comparisons); score 2 was compared with scores 3, 4 and 5 (three comparisons); score 3 was compared with scores 4 and 5 (two comparisons); and score 4 was compared with score 5 (one comparison).) In this case the areas under two ROC curves were compared according to the correction for paired samples proposed by Hanley and McNeill[18]. We then applied Bonferroni's correction, considering a difference between two ROC curves represented by a p value of < 0.005 as significant. The same methodology, without the correction for paired samples, was used to compare the ROC curves obtained in the three centers with the new score.

The diagnostic accuracy (sum of the true positives and true negatives expressed as a percentage of the total number of cases) of each morphological score, according to the best cut-off value observed by ROC curves in the whole series, was then calculated for (1) the whole series; (2) lesions of ≤ 5 cm; and (3) borderline lesions and stage I cancers. For each score, both the cut-off level that maximised the sum of the sensitivity and specificity and the 'best' clinical cut-off level are reported. The 'best' clinical cut-off level was chosen according to the results of the ROC curves and the general criteria of a screening test. The differences in diagnostic accuracy between the multicenter score and the other four systems were tested using McNemar's test.

RESULTS

Two hundred and sixty-one masses proved to be benign: among these we correctly identified 11 tubal post-infective lesions and eight paraovarian masses. Sixty-nine neoplasms were malignant; two tubal cancers which occurred in this group were thought to be of ovarian origin. The prevalence of malignant lesions in the three centers was 20% (center 1), 19% (center 2) and 25% (center 3). No significant difference was found between the ROC curves obtained in the three centers with the new score.

Table 3 illustrates the dimensions of malignant and benign lesions according to their pathological origin. Thirty malignant lesions were limited to the ovaries (19 borderline and 11 stage I). In 171 benign and 36 malignant lesions (63%) the mean diameter was less than 5 cm.

Figure 1 shows ROC curves for the five scoring systems under the condition that typical dermoids were given the lowest values for each score, except for Lerner's score. This was the case for 41 out of 51 pathologically proven cystic teratomas. The areas under the ROC curves ranged from 0.75 to 0.84%. The only significant difference was found between the curve of the multicenter scores and the curve of Lerner's scores ($p < 0.0004$). When the ROC curves of Granberg, Sassone and De Priest, according to their original description (which did not take into account the possible false-positive data caused by typical cystic teratomas), were compared with the multicenter scores, a significantly smaller area under the ROC curve was found for all these three scoring systems ($p < 0.0001$) (Figure 2).

The diagnostic accuracy of the sonographic test for the whole series is reported in Table 4. This table shows both the values achieved by the cut-off level that maximized the sum of the sensitivity and specificity according to the ROC curves (Figure 2, scoring systems not corrected for dermoid cysts) and the values reached by a cut-off level that allowed

Table 3 Histological diagnosis, number of cases and diameter (mean ± SD, where applicable) of adnexal neoplasms in the present series

Histological diagnosis	*n*	*Mean diameter* (mm)
Ovarian carcinoma, borderline and stage I	30	63 ± 34
Ovarian carcinoma, stages II, III and IV	35	63 ± 33
Tubal cancer, stage I	1	50
Tubal cancer, stage III	1	45
Ovarian lymphoma	1	50
Krukenberg tumor	1	48
Total malignant lesions	69	63 ± 33
Cystadenoma	71	58 ± 34
Serous cyst	16	53 ± 27
Cystic teratoma	51	52 ± 25
Endometrioma	69	44 ± 16
Fibroid	17	46 ± 28
Hemorrhagic corpus luteum	17	41 ± 10
Paraovarian cyst	8	39 ± 13
Tubal post-infection lesion	11	52 ± 20
Peritoneal pseudocyst	1	60
Total benign lesions	261	50 ± 26

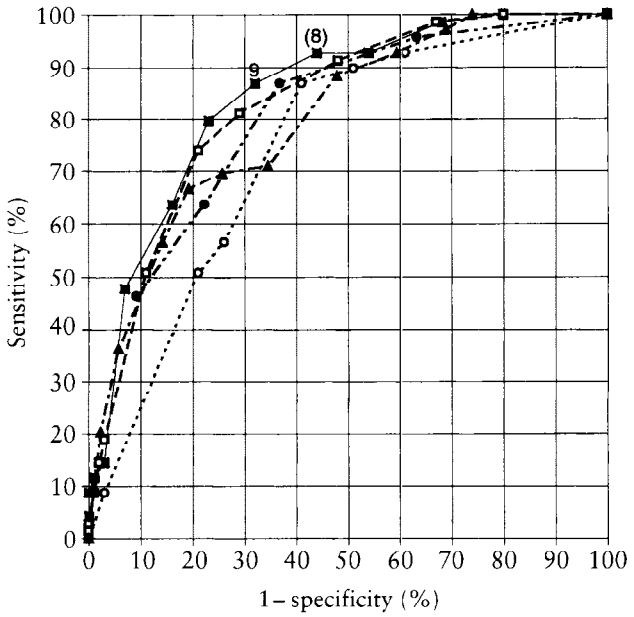

Figure 1 Receiver operator characteristic (ROC) curves calculated for the whole series with the correction for dermoid cysts (except for Lerner's score): multicenter curve (full square): area under the ROC curve 0.84, SE 0.02; curve of Sassone and colleagues[5] (open square): area under the ROC curve 0.82, SE 0.03; curve of De Priest and colleagues[6] (full triangle): area under the ROC curve 0.81, SE 0.03; curve of Granberg and colleagues[2] (full circle): area under the ROC curve 0.80, SE 0.03; curve of Lerner and colleagues[7] (open circle): area under the ROC curve 0.75, SE 0.03. For the multicenter score, both the cut-off value that maximized the sum of the sensitivity and specificity according to the ROC curves and the cut-off value that allowed enhancement of sensitivity (between parentheses) are reported

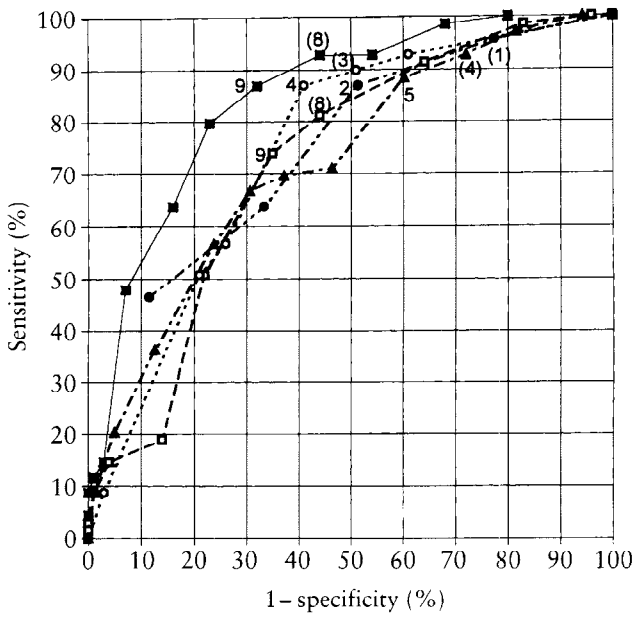

Figure 2 Receiver operator characteristic (ROC) curves calculated for the whole series according to the original description of each score: multicenter curve (full square): area under the ROC curve 0.84, SE 0.02; curve of Sassone and colleagues[5] (open square): area under the ROC curve 0.72, SE 0.03; curve of De Priest and colleagues[6] (full triangle): area under the ROC curve 0.73, SE 0.03; curve of Granberg and colleagues[2] (full circle): area under the ROC curve 0.74, SE 0.03; curve of Lerner and colleagues[7] (open circle): area under the curve 0.75, SE 0.03. For each score both the cut-off value that maximized the sum of the sensitivity and specificity according to the ROC curves and the cut-off value that allowed enhancement of sensitivity (between parentheses) are reported

Table 4 Diagnostic accuracy of the five scoring systems

Scoring systems	*Score*	*Sensitivity* (%)	*Specificity* (%)	*Positive predictive value* (%)	*Negative predictive value* (%)
Multicenter	9	87	67	41	95
	(8)	93	56	36	97
Granberg	2	87	49	31	93
et al.	(1)	96	23	25	95
Sassone	9*	74	65	36	90
et al.	(8)	81	56	33	92
De Priest	5*	88	40	28	93
et al.	(4)	93	28	25	94
Lerner	4	87	59	36	95
et al.	(3)*	90	49	32	95

Score without parentheses is the cut-off value that maximizes the sum of sensitivity and specificity according to the ROC curves obtained in the present series; score between parentheses is the cut-off value that allows maximum sensitivity without a dramatic fall in specificity; *value reported in the original paper

increased sensitivity. The multicenter, Sassone and Lerner scores allowed maximum sensitivity without a dramatic fall in specificity.

Table 5 Comparison of the diagnostic accuracy of the multicenter score vs. the other four scores (by McNemar's test, significant: $p < 0.004$)

	Multicenter ≥ 9	*Granberg* ≥ 2	*Sassone* ≥ 9	*De Priest* ≥ 5	*Lerner* ≥ 4
Whole series					
Accuracy	72%	57%	67%	50%	65%
p Value		0.0001	0.092	0.0001	0.077
Lesions ≤ 5 cm					
Accuracy	69%	53%	63%	52%	63%
p Value		0.0002	0.349	0.0001	0.142
Cancers ≤ stage I					
Accuracy	69%	52%	65%	44%	62%
p Value		0.0001	0.176	0.0001	0.069

Table 5 shows the results of McNemar's test, which compared the accuracy of the multicenter scoring system against that of the other four systems, for the whole series, for lesions of ≤ 5 cm and for cancers of ≤ stage I.

The diagnostic accuracy of the multicenter scoring system achieved by the 'best' clinical cut-off value (score ≥ 8) for lesions of ≤ 5 cm was: sensitivity 92%, specificity 53%, positive and negative predictive values 29% and 97%, respectively.

The diagnostic accuracy of the multicenter scoring system achieved by the 'best' clinical cut-off value (score ≥ 8) for cancers of ≤ stage I was: sensitivity 87%, specificity 56%, positive and negative predictive values 19% and 97%, respectively.

DISCUSSION

Morphological scoring systems allow us to overcome subjective interpretation of ovarian lesions, to compare the results of different centers and to assess a precise role for this diagnostic technique in the work-up of patients with persistent ovarian lesions. Scoring systems do have limitations, however. A 'score' tells us that the system we are using to investigate a problem is imprecise. This is the case with ovarian neoplasms and sonographic imaging. Although the image resolution is actually below 1 mm, we are not using a microscope, and we are limited to a gross, indirect description.

The aim of developing a new scoring system was to improve the diagnostic capability by incorporating systematic descriptive indices into a morphology score. The opinion was that the indices of Granberg and co-workers[2], De Priest and co-workers[6] and Lerner and co-workers[7] did not satisfy the criteria of a systematic description of the lesion, and that variable efforts are required to fit each characteristic of the lesion into these scoring systems, thus again relying upon the subjective experience of the individual sonologist. The score of Sassone and colleagues[5] has the advantage of being more descriptive and helpful for routine examination but, in our opinion, suffers from inconsistencies, including the fact that wall thickness *per se* probably does not add much sensitivity, and the additional sensitivity provided by papillarities is not properly emphasized. An important achievement of the scoring system of Lerner and co-workers[7] was that, by means of a multiple regression analysis, it assigned to various aspects of ovarian masses a non-continuous scoring value.

Another practical limitation of all scoring systems is that the typical dermoid cyst is given high scores even though it is often the most easily identified by sonologists. The scoring system of Lerner and colleagues[7] attempts to overcome this pitfall by assigning a value of 0 to the typical sonographic shadowing produced by the presence of a dermoid and a value of 1 to any lesions without shadowing.

The guidelines of our scoring system attempt to overcome these limitations by: (1) having an easy-to-use flow chart to describe the lesion; (2) assigning a non-continuous point value to the morphological aspects of the lesion; (3) giving the typical content of the endohemorrhagic corpus luteum the same value as that of anechogenic content; and (4) excluding typical dermoids from detailed scoring. Dermoids are treated as they are in daily clinical practice. When the lesion is typical (41 out of 51 in the present series) the scoring system does not apply, a pathological diagnosis is made and the lowest score is given, to allow for a retrospective estimate of the accuracy of the technique.

The results obtained in the three centers with this new scoring system were not significantly different. The ROC curves showed that the multicenter score proved to be significantly better than the scores of Sassone[5], De Priest[6] and Granberg[2] ($p < 0.0001$), and Lerner[7] ($p < 0.004$). However, when the correction for typical dermoids was applied to the curves of Granberg[2], Sassone[5] and De Priest[6], the results showed that the multicenter score was relatively but not statistically significantly better than the other three scores. This correction was not applied for Lerner's score[7], which was designed to account for this limitation.

These results were also partially confirmed by the comparison of the accuracy according to McNemar's tests, which showed a significantly better performance of the multicenter scores vs. the scores of Granberg and De Priest. This significantly better performance was confirmed by the McNemar's test also for 'small' lesions and 'early' cancer (≤ stage I).

It is of great interest to see how, in spite of different designs, the five scoring systems performed very closely with regards to sensitivity in diagnosing malignant masses. Specificity was much improved by our scoring system as a result of the correction for typical dermoids and for typical endohemorrhagic corpora lutea. However, as often happens, the results of a test do not always conform to those of the original pilot trial. Granberg[2], Sassone[5], De Priest[6] and Lerner[7] and their colleagues reported a sensitivity of 82%, 100%, 100% and 97%, respectively. These four scoring systems, applied to the present series, achieved a sensitivity of 87%, 74%, 88% and 87%, respectively. The number of malignant lesions in our series was twice that of Lerner[7] and Granberg[2], and more than three times that of Sassone[5] and De Priest[6]. The pleiomorphism of the malignant lesion is better represented in larger series and could explain the consistently lower sensitivity we observed.

The best cut-off value of our system that maximized the sum of sensitivity and specificity according to the ROC curves obtained on the present series was 9. This value achieved a sensitivity of 87% and a specificity of 67%. However, since the clinical cost of false-normal and false-abnormal tests is not the same, we were not interested in this mathematical value. For the clinical purposes of this screening test we also checked a possible cut-off value of 8, which increased the sensitivity to 93% and decreased the specificity to 56%. With the same criteria used for the scores of the different authors, the following values were obtained for sensitivity: 96%, 81%, 93% and 90%; and for specificity 23%, 56%, 28% and 49%, respectively. The specificity observed for Granberg and De Priest is probably too low to be acceptable, even for a non-invasive screening test such as transvaginal sonography.

A similar diagnostic accuracy was achieved by a score of ≤ 8 for masses ≤ 5 cm in mean diameter. The sensitivity was 92% and the specificity was 53%. Again, to our surprise, the results of the scoring systems did not worsen when the sensitivity was checked against lesions of borderline malignancy and stage I lesions only (sensitivity 87%, specificity 56%).

To the best of our knowledge, these two specific aspects of the diagnostic accuracy of morphological scoring following transvaginal sonography, i.e. small lesions and

'early' cancer, have not been considered of particular importance in previously reported series. In our opinion, these areas of non-invasive diagnosis are of paramount importance. The diagnosis of small, frequently non-palpable masses relies solely on sonographic imaging. Similarly, 'early' cancer can probably benefit from more timely diagnosis and therapy than 'late' cancer, and direct proof of the feasibility of early diagnosis strengthens the clinical role of transvaginal sonography in screening for ovarian cancer.

In conclusion, the five different scoring systems showed very high and similar diagnostic accuracy in distinguishing malignant from benign lesions – in large masses, in masses of ≤ 5 cm and in early cancer. The new score adopted in this multicenter study proved to be better than the other four traditional scores. This was mainly due to the new criteria introduced for typical dermoids and endohemorrhagic corpora lutea. As expected, a completely reliable differentiation of malignant masses cannot be obtained by sonographic imaging alone.

ACKNOWLEDGEMENTS

AIRCS, the Italian Association for Cancer Research, has supported this study. Our sincere thanks go to Dr Walter Torri, form the 'Mario Negri' Research Institute of Milan, for his careful statistical advice.

REFERENCES

1. Di Saia, P. J. and Creasman, W. T. (1989). *Clinical Gynecologic Oncology*, 3rd edn. (St. Louis: Mosby)
2. Granberg, S., Nosrom, A. and Wikland, M. (1990). Tumors in the lower pelvis as imaged by vaginal sonography. *Gynecol. Oncol.*, **37**, 224–9
3. Van Nagell, J. R., Higgins, R. V., Donaldson, E. S., Gallion, H. H., Powell, D. E., Pavlik, E. J., Woods, C. H. and Thompson, E. A. (1990). Transvaginal sonography as a screening method for ovarian cancer: a report of the first 1000 cases screened. *Cancer*, **65**, 573–7
4. Bourne, T. H., Campbell, S., Reynolds, K. M., Whitehead, M. I., Hampson, J., Royston, P., Crayford, T. J. B. and Collins, W. P. (1993). Screening for early familial ovarian cancer with transvaginal ultrasonography and color blood flow imaging. *Br. Med. J.*, **306**, 1025–9
5. Sassone, A. M., Timor-Tritsch, I. E., Artner, A., Carolyn, W. and Warren, W. B. (1991). Transvaginal sonographic characterization of ovarian disease: evaluation of a new scoring system to predict ovarian malignancy. *Obstet. Gynecol.*, **78**, 70–6
6. De Priest, P. D., Shenson, D., Fried, A., Hunter, J. E., Andrews, S. J., Gallion, H. H., Pavlik, E. J., Kryscio, R. J. and Van Nagell, J. R. (1993). A morphology index based on sonographic findings in ovarian cancer. *Gynecol. Oncol.*, **51**, 7–11
7. Lerner, J. P., Timor-Tritsch, I. E., Federman, A. and Abramovich, G. (1994). Transvaginal ultrasonographic characterization of ovarian masses with an improved, weighted scoring system. *Am. J. Obstet. Gynecol.*, **170**, 81–5
8. Bromley, B., Goodman, H. and Benacerraf, B. R. (1994). Comparison between sonographic morphology and doppler waveform for the diagnosis of ovarian malignancy. *Obstet. Gynecol.*, **83**, 434–7
9. Bourne, T. H., Campbell, S., Steer, C., Whitehead, M. I. and Collins, W. P. (1989). Transvaginal colour flow imaging: a possible new screening technique for ovarian cancer. *Br. Med. J.*, **299**, 1367–70
10. Zanetta, G., Vergani, P. and Lissoni, A. (1994). Color doppler ultrasound in the preoperative assessment of adnexal masses. *Acta Obstet. Gynecol. Scand.*, **73**, 637–41
11. Kurjak, A., Zalud, I. and Alfirevic, Z. (1991). Evaluation of adnexal masses with transvaginal color ultrasound. *J. Ultrasound Med.*, **10**, 295–9
12. Fleisher, A. C., Rogers, W. H., Rao, B. K., Kepple, D. M. and Jones, H. W. (1991). Assessment of ovarian tumor vascularity with transvaginal color doppler sonography. *J. Ultrasound Med.*, **10**, 563–8
13. Hata, K., Makihara, K., Hata, T., Takahashi, K. and Kitao, M. (1991). Transvaginal color doppler imaging for hemodynamic assessment of reproductive tract tumors. *Int. J. Gynecol. Obstet.*, **36**, 301–8
14. Valentin, L., Sladkevicius, P. and Marsàl, K. (1994). Limited contribution of doppler velocimetry to the differential diagnosis of extrauterine pelvic tumors. *Obstet. Gynecol.*, **83**, 425–33
15. Bourne, T. H. (1994). Should clinical decisions be made about ovarian masses using transvaginal color Doppler? *Ultrasound Obstet. Gynecol.*, **4**, 357–60
16. Buy, J. N., Ghossain, M. A., Hugol, D., Hassen, K., Sciot, C., Truc, J. B., Poitout, P. and Vadrot, D. (1996). Characterization of adnexal masses: combination of color Doppler and conventional sonography compared with spectral Doppler analysis alone and conventional sonography alone. *Am. J. Roentgenol.*, **166**, 385–93
17. Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36
18. Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–43