Review

Colposcopy for the diagnosis of squamous intraepithelial lesions: A meta-analysis

Michele Follen Mitchell, MD, MS,
David Schottenfeld, MD, MS,
Guillermo Tortolero-Luna, MD, PhD,
Scott B. Cantor, PhD, and
Rebecca Richards-Kortum, PhD

Objective: **To quantify by meta-analysis the performance of colposcopy to set a standard against which new technologies can be compared.**

Data Sources: **MEDLINE was searched for articles on colposcopy for diagnosis of squamous intraepithelial lesions (SIL). The search selected articles from 1960 to 1996 combining the key word "colposcopy" with key words "diagnosis," "positive predictive value," "negative predictive value," "likelihood ratio," and "receiver operating characteristic (ROC) curve."**

Methods of Study Selection: **Articles were selected if the authors studied a population of patients with abnormal screening Papanicolaou smears and presented raw data showing for each cervical lesion type the number of patients judged positive and negative by colposcopic impression versus the standard of colposcopic biopsy results. Nine of 86 studies met these criteria.**

Tabulation, Integration, and Results: **Biopsies had been categorized as normal, atypia, cervical intraepithelial neoplasia (CIN) I, CIN II, CIN III, carcinoma in situ, and invasive cancer; we recalculated performance measures using the Bethesda system. Overall sensitivity, specificity, likelihood ratios, ROC curves, and the corresponding areas under the curves were calculated. The average weighted sensitivity of diagnostic colposcopy for the threshold normal compared with all cervix abnormalities (atypia, low-grade SIL, high-grade SIL, cancer) was 96% and the average weighted specificity 48%. For the threshold normal cervix and low-grade SIL compared with high-grade SIL and cancer, average weighted sensitivity was 85% and average weighted specificity 69%. Likelihood ratios generated small but important changes in probability for distinguishing normal cervix and low-grade SIL from high-grade SIL and cancer. Areas under the ROC curve were 0.80 for the threshold normal cervix compared with all abnormalities and 0.82 for the threshold normal cervix and low-grade SIL compared with high-grade SIL and cancer.**

Conclusion: **Colposcopy compares favorably with other medical diagnostic tests in terms of sensitivity, specificity, and area under the ROC curve. New diagnostic methods for the cervix can be compared with colposcopy using these quantified values. (Obstet Gynecol 1998;91:626–31. © 1998 by The American College of Obstetricians and Gynecologists.)**

The standard of care for patients with abnormal Papanicolaou smear results is to perform colposcopy and directed biopsy. The National Cancer Institute estimates that 50 million Papanicolaou smears are performed annually in the United States; of these, approximately 2.5 million (5%) show evidence of low-grade abnormalities.[1] The cost of colposcopic evaluations and interventions for low-grade lesions alone approaches $6 billion annually in the United States.[1]

Analysis of the cost-effectiveness of cervical cancer screening programs has shown that the rate of cost increase is most heavily influenced by the false-positive rate; for example, in simulation studies assuming annual screening of elderly women, more than half of the cost of screening and detection programs was devoted to following up false-positive results.[2] A number of new clinical management strategies and technologies have been proposed and tested to address the need to improve screening and detection of squamous intraepithelial lesions (SIL) and cervical cancer. These include human papillomavirus (HPV) testing, cervicography, speculoscopy, the polar probe, and fluorescence spectroscopy. Some of these have been proposed to supplement colposcopy as a diagnostic tool; others have been proposed to replace colposcopy.

In evaluating these new technologies to determine their optimal clinical role, their performance must be compared with the "gold" standard—biopsy—and with the standard of care—colposcopy. A quantitative means of comparing these technologies was needed. Therefore, in this study, we performed a meta-analysis of results reported in the literature and computed an overall sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratio, receiver operating characteristic (ROC) curve, and area under the ROC curve for colposcopy. Our objective was to quan-

From the Department of Gynecologic Oncology and the Section of General Internal Medicine, Department of Medical Specialties, University of Texas M. D. Anderson Cancer Center, Houston, Texas; the Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan; and the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Texas.

The authors gratefully acknowledge financial support from The Whitaker Foundation.

tify the performance of colposcopy to obtain a standard against which new technologies can be compared easily.

Data Sources

We searched the printed *Cumulative Index Medicus* for articles published from 1960 to 1965 and the MEDLINE database for articles published from 1966 to 1996. The key word "colposcopy" was used in conjunction with key words "diagnosis," "positive predictive value," "negative predictive value," "likelihood ratio," and "receiver operating characteristic curve." The search was limited to articles published in English.

Methods of Study Selection

Eighty-six articles were identified by the MEDLINE search. Articles were included in the meta-analysis if they met two criteria. First, colposcopy had to be used for diagnosis; in other words, women had been referred to a colposcopy clinic because of abnormal Papanicolaou smear results. (Studies were excluded if referrals were based on abnormal bleeding.) The studies therefore had high-prevalence populations. Second, the article had to include raw data in the form of a table comparing colposcopic impression with results of colposcopically directed biopsy, broken down by disease categories (normal cervix, atypia, HPV, cervical intraepithelial neoplasia (CIN) I, CIN II, CIN III, carcinoma in situ (CIS), and invasive cancer). In other words, studies were selected if they included a table comparing the accuracy of the colposcopy results with the "gold" standard of biopsy results. Raw data were needed to calculate overall summary statistics and analyze different combinations of disease categories (we used two different thresholds: one distinguishing normal tissue from all abnormalities and one distinguishing normal tissue, atypia, and low-grade SIL from high-grade SIL and cancer).

Tabulation and Integration

Patients with biopsy results demonstrating no abnormality or atypia were considered "disease absent," and patients with biopsies showing HPV, CIN I, CIN II, CIN III, CIS, or invasive cancer were considered "disease present." The studies predated the Bethesda classification[1]; to adapt the data to the Bethesda system, we combined HPV and CIN I into the category "low-grade SIL" and CIN II, CIN III, and CIS into "high-grade SIL."

Nine studies[3–11] met the criteria for meta-analysis using the threshold of normal cervix compared with all abnormalities. Eight of these studies[3–7,9–11] also were eligible for meta-analysis using the threshold of normal cervix, atypia, and low-grade SIL compared with high-grade SIL and cancer; the ninth study[8] did not separate data for CIN I and CIN II, so low-grade SIL and high-grade SIL could not be distinguished. Using these two thresholds, sensitivity and specificity were calculated.[12–16] Weighting by sample size, weighted mean sensitivity and specificity also were determined. Positive and negative predictive values were calculated using Bayes' theorem.

We also calculated likelihood ratios, which compare the pretest probability of a positive or negative test when disease is either present or absent. To calculate these ratios, the study population was divided into those with disease present and those with disease absent. The colposcopic impression was stratified as normal, atypia, low-grade SIL, high-grade SIL, or cancer; there was an insufficient number of patients in the cancer category to allow the stratification of cancer as a separate diagnosis. The likelihood of a positive test was expressed as a fraction (ratio); the numerator was the proportion of patients within the diagnostic category who were positive for disease, and the denominator was the proportion of patients within each diagnostic category who were negative for disease ([sensitivity/false-positives in patients with disease present]/[sensitivity/false-positives in patients with disease absent]). Likelihood ratios less than 0.1 and greater than 10 generate large and often conclusive changes from pretest to post-test probability. Those from 0.1 to 0.2 and 5 to 10 generate moderate shifts, those from 0.2 to 0.5 and 2 to 5 generate small shifts, and those from 0.5 to 1.0 and 1.0 to 2.0 generate negligible shifts.[15,16]

We used meta-analysis to estimate a summary ROC curve from these independent reports of sensitivity and specificity of a test.[12–14,17] We plotted sensitivity and 100− specificity from each of the nine studies using the two thresholds described above. From these data points, we calculated a summary ROC curve, which suggests an overall performance of the diagnostic test across the spectrum of diagnostic settings reported in the literature. In this analysis, the horizontal and vertical axes of the plot are transformed in such a way that the data can be reasonably described by a straight line.[12] The goal is to estimate a smooth curve that passes through, or near, all independent data points.[12] We used the logistic transform method recommended by Littenberg and Moses.[12] The mathematical details[12] and the properties[17] of the transform have been described in detail. We also calculated the area under the ROC curve, which yields a value that can be used easily to compare the performance of diagnostic tests.

Table 1. Measures of Effectiveness of Colposcopy for Distinguishing Normal Tissue From Abnormal Tissue (Atypia, Low-Grade Squamous Intraepithelial Lesions, High-Grade Squamous Intraepithelial Lesions, Cancer)

Study	Number of patients	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)
Benedet et al[3]	549	.99 (±.06)	.53 (±.04)	.95 (±.02)	.93 (±.02)
Benedet et al[4]	3252	.95 (±.07)	.44 (±.02)	.83 (±.01)	.74 (±.02)
Cristoforoni et al[5]	188	.98 (±.02)	.34 (±.07)	.77 (±.06)	.88 (±.05)
Edebiri[6]	222	.87 (±.04)	.67 (±.06)	.80 (±.05)	.77 (±.06)
Ferris and Miller[7]	205	.97 (±.02)	.23 (±.06)	.90 (±.04)	.52 (±.07)
Javaheri and Fejgin[8]	903	.99 (±.06)	.87 (±.01)	.96 (±.01)	.99 (±.06)
Lozowski et al[9]	151	.96 (±.03)	.29 (±.07)	.80 (±.06)	.71 (±.07)
Seshadri et al[10]	152	.87 (±.05)	.34 (±.07)	.53 (±.08)	.72 (±.07)
Stafl and Mattingly[11]	659	.99 (±.03)	.26 (±.03)	.81 (±.03)	.87 (±.03)
Mean		.95 (±.01)	.45 (±.01)	.82 (±.01)	.79 (±.01)
Weighted mean		.96 (±.01)	.48 (±.01)		

CI = confidence interval.

Results

Among the nine studies[3–11] identified, for distinguishing normal cervix from all other diagnoses, the individual estimations of sensitivity of diagnostic colposcopy (87–99%) were high, whereas those of specificity (23–87%) were lower (Table 1). Similarly, among the eight studies[3–7,9–11] with fully separated disease categories, for distinguishing normal cervix, atypia, and low-grade SIL from high-grade SIL and cancer, the estimations of sensitivity of diagnostic colposcopy (64–99%) were high, whereas those of specificity (30–93%) were lower (Table 2). Mean weighted specificity improved slightly (69% versus 48%) using the latter threshold. High-grade lesions appeared to have distinguishing characteristics that allowed them to be better separated from low-grade lesions than was possible for separating low-grade lesions from normal cervix.

The likelihood ratios (Table 3) show a clear demarcation among diagnoses with small but consistent and important shifts occurring between normal cervix, atypia, or low-grade SIL and high-grade SIL or cancerous lesions. This information supports that presented for the distribution of proportional sensitivities and specificities, indicating that clinically there is a larger appreciable difference when distinguishing high-grade SIL and cancer from less-serious diagnoses than when distinguishing low-grade SIL from normal cervix and inflammation. This difference supports the validity of the Bethesda classification.[1]

The weighted ROC curves (Figures 1 and 2) show reasonable performance for diagnostic colposcopy using both thresholds, with areas under the curves of 0.80 for distinguishing normal cervix from all abnormalities and 0.82 for distinguishing normal cervix, atypia, and low-grade SIL from high-grade SIL and cancer. The area under the ROC curve for colposcopy compares favorably with those of other diagnostic tests used in medicine (Figure 3).

Table 2. Measures of Effectiveness of Colposcopy for Distinguishing Normal Tissue, Atypia, and Low-Grade Squamous Intraepithelial Lesions From High-Grade Squamous Intraepithelial Lesions and Cancer

Study	Number of patients	Sensitivity (95% CI)	Specificity (95% CI)	Positive predictive value (95% CI)	Negative predictive value (95% CI)
Benedet et al[3]	549	.95 (±.02)	.64 (±.04)	.84 (±.03)	.88 (±.03)
Benedet et al[4]	3252	.72 (±.02)	.69 (±.02)	.72 (±.02)	.68 (±.02)
Cristoforoni et al[5]	188	.64 (±.07)	.92 (±.04)	.20 (±.06)	.92 (±.04)
Edebiri[6]	222	.80 (±.05)	.66 (±.06)	.70 (±.06)	.77 (±.06)
Ferris and Miller[7]	205	.30 (±.06)	.93 (±.04)	.53 (±.07)	.83 (±.05)
Lozowski et al[9]	151	.96 (±.03)	.56 (±.08)	.75 (±.07)	.92 (±.04)
Seshadri et al[10]	152	.87 (±.05)	.39 (±.07)	.44 (±.08)	.83 (±.06)
Stafl and Mattingly[11]	659	.99 (±.04)	.59 (±.04)	.38 (±.04)	.97 (±.02)
Mean		.79 (±.01)	.67 (±.01)	.57 (±.01)	.85 (±.01)
Weighted mean		.85 (±.01)	.69 (±.01)		

CI = confidence interval.

Table 3. Likelihood Ratios of Colposcopy for Identifying Normal Tissue, Atypia, Low-Grade and High-Grade Squamous Intraepithelial Lesions, and Cancer

Study	Diagnosis	Number of patients	Proportion biopsy positive	Proportion biopsy negative	Likelihood ratio
Benedet et al[3]		549			
	Negative		2/436	60/113	0.01
	Low-grade SIL		41/436	28/113	0.38
	High-grade SIL or cancer		393/436	25/113	4.1
Benedet et al[4]		3252			
	Negative		131/2415	370/837	0.12
	Low-grade SIL		613/2415	251/837	0.83
	High-grade SIL		1671/2415	216/837	2.65
Cristoforoni et al[5]		188			
	Negative		3/130	20/58	0.07
	Low-grade SIL		96/130	36/58	1.19
	High-grade SIL		31/130	2/58	8.0
Edebiri[6]		222			
	Negative		17/130	62/92	0.19
	Low-grade SIL		12/130	13/92	0.65
	High-grade SIL or cancer		101/130	17/92	4.22
Ferris and Miller[7]		205			
	Negative		3/115	21/90	0.11
	Low-grade SIL		91/115	65/90	1.10
	High-grade SIL		21/115	4/90	4.16
Javaheri and Fejgin[8]		903			
	Negative		1/681	194/222	0.002
	CIN I–II		299/681	21/222	4.6
	CIN III, CIS, cancer		381/681	7/222	18.7
Lozowski et al[9]		151			
	Atypia		4/113	11/38	0.12
	Low-grade SIL		19/113	16/38	0.40
	High-grade SIL or cancer		90/113	11/38	2.76
Seshadri et al[10]		152			
	Negative		9/70	28/82	0.38
	Low-grade SIL		13/70	36/82	0.43
	High-grade SIL or cancer		48/70	18/82	2.91
Stafl and Mattingly[11]		659			
	Negative		6/499	42/160	0.05
	Low-grade SIL		97/499	66/160	0.46
	High-grade SIL or cancer		396/499	52/160	2.49

SIL = squamous intraepithelial lesion; CIN = cervical intraepithelial neoplasia; CIS = carcinoma in situ.

Discussion

Colposcopy is a visual technique that requires extensive training and experience. A suitable volume of patients must be seen yearly in order to maintain these skills.[18] Another disadvantage of current technologies is that the patient and care provider must wait 1–2 weeks for the results of the Papanicolaou smear, endocervical curettage, and biopsy. Technology that eliminates the need for training or reduces the delay for diagnostic results could improve patient care appreciably if the new technology performs as well as colposcopy.

This article presents methods used for the evaluation of diagnostic and screening tests in medicine to set a standard for the evaluation of new diagnostic technologies in comparison with colposcopy. Using three methods, we quantified the detection capabilities of the colposcopic technique; these findings can be used as a basis of comparison for emerging experimental technologies, such as fluorescence spectroscopy and the polar probe, and for technologies being tested, such as cervicography, HPV testing, and speculoscopy.

The high sensitivity and low specificity of colposcopy are most likely due to the "overcalling" of low-grade lesions. This explanation is confirmed by the finding that the specificity improved when the threshold was set to distinguish high-grade lesions and cancer from lesser abnormalities. Furthermore, the likelihood ratios showed much larger shifts between low-grade and high-grade lesions than between normal cervix and low-grade lesions. This distinction results partially from

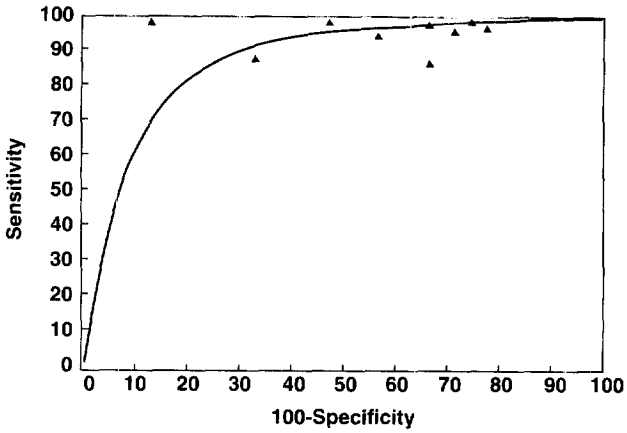

Figure 1. Sensitivity and 100-specificity reported for diagnostic colposcopic impression in nine studies (each represented by a point in the graph) relative to the standard of colposcopically directed biopsy for distinguishing a normal or inflamed cervix from all other abnormalities (atypia, low- or high-grade squamous intraepithelial lesion, or cancer). The *solid line* shows the receiver operating characteristic curve determined from regression analysis.

the fact that in the classification of colposcopic lesions, vascular atypia is the hallmark of higher grade lesions, yet vascular atypia can be the result of an HPV infection without associated SIL. Similarly, inflammation sometimes can result in vascular atypia that is worrisome.[19]

The ROC curve is dependent on the composition of the tested population either with or without preneoplastic disease, as well as on varying sensitivities and specificities. As a result, errors can occur if summary ROC curves are constructed from populations not resembling those under study, and further analysis and

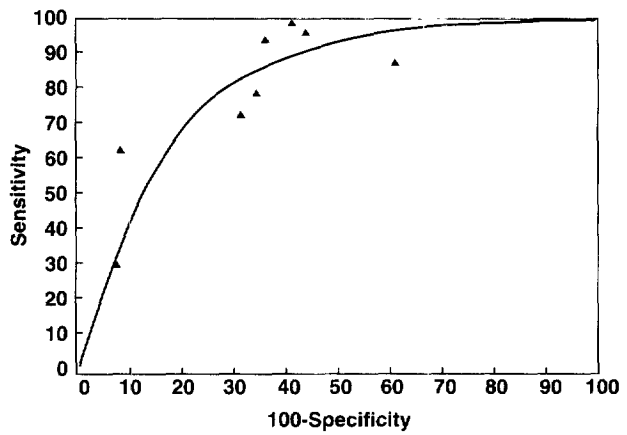

Figure 2. Sensitivity and 100-specificity reported for diagnostic colposcopic impression in eight studies (each represented by a point in the graph) relative to the standard of colposcopically directed biopsy for distinguishing a normal cervix and low-grade squamous intraepithelial lesions from high-grade squamous intraepithelial lesions and cancer. The *solid line* shows the receiver operating characteristic curve determined from regression analysis.

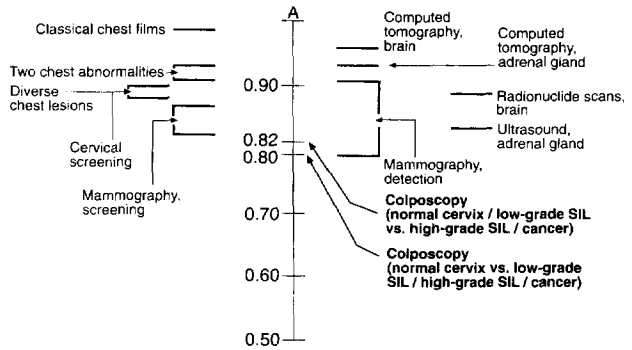

Figure 3. Comparison of areas under the receiver operating characteristic curves of many diagnostic tests in medicine. A = area under the curve; SIL = squamous intraepithelial lesion. (Adapted with permission from Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285–93.)

assessment of test performance will be required in a particular clinical setting that reflects the population of interest. All of the studies selected were performed by experienced colposcopists who run large colposcopy clinics. Experienced practitioners are more likely than inexperienced practitioners to publish their experience, and their performance is probably better than that of less experienced practitioners.

The ROC curve compensates for the wide variation in the sensitivity and specificity of colposcopy among the reported studies. This variation, which is typical of studies on medical tests, could have arisen from three principal factors: 1) tradeoffs between sensitivity and specificity as the threshold for positive diagnosis is varied, 2) differences in the tested populations, and 3) differences in the skill of the colposcopists. The method of Littenberg and Moses[12] takes these changes into account in the construction of the ROC curve. As discussed by Littenberg and Moses,[12] the method used here to estimate the ROC curves summarizes many reports, without specifying which variables differ from study to study. The method can be used to recognize outliers by comparing the actual data points to the estimated ROC curve. All the individual data points fell near our estimated ROC curves (Figures 1 and 2).

This study attempts to quantify the performance of colposcopy for diagnosis. New technologies must add value by improving performance or lowering cost. These results should facilitate the comparison of the performance of new diagnostic tests with that of colposcopy.

References

1. Kurman RJ, Henson DE, Herbst AL, Noller KL, Schiffman MH. Interim guidelines for management of abnormal cervical cytology. JAMA 1994;271:1866–9.

2. Fahs MC, Mandelblatt J, Schechter C, Muller C. Cost effectiveness of cervical cancer screening for the elderly. Ann Intern Med 1992;117:520–7.
3. Benedet JL, Boyes DA, Nichols TM, Millner A. Colposcopic evaluation of patients with abnormal cervical cytology. Br J Obstet Gynaecol 1976;83:177–82.
4. Benedet JL, Anderson GH, Matsic JP, Miller DM. A quality-control program for colposcopic practice. Obstet Gynecol 1991;78:872–5.
5. Cristoforoni PM, Gerbaldo D, Perino A. Piccoli R, Montz FJ, Capitanio GL. Computerized colposcopy: Results of a pilot study and analysis of its clinical relevance. Obstet Gynecol 1995;85: 1011–6.
6. Edebiri AA. The relative significance of colposcopic descriptive appearances in the diagnosis of cervical intraepithelial neoplasia. Int J Gynaecol Obstet 1990;33:23–9.
7. Ferris DG, Miller MD. Colposcopic accuracy in a residency training program: Defining competency and proficiency. J Fam Pract 1993; 36:515–20.
8. Javaheri G, Fejgin MD. Diagnostic value of colposcopy in the investigation of cervical neoplasia. Am J Obstet Gynecol 1980;137: 588–94.
9. Lozowski MS, Mishriki Y, Talebian F, Solitare G. The combined use of cytology and colposcopy in enhancing diagnostic accuracy in preclinical lesions of the uterine cervix. Acta Cytol 1982;26:285–91.
10. Seshadri L, Jairaj P, Krishnaswami H. Colposcopy in the diagnosis of cervical neoplasia. Indian J Cancer 1990;27:180–6.
11. Stafl A, Mattingly RF. Colposcopic diagnosis of cervical neoplasia. Obstet Gynecol 1973;41:168–76.
12. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method. Med Decis Making 1993;13:313–21.
13. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.
14. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285–93.
15. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? JAMA 1994;271:389–91.
16. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? JAMA 1994;271:703–7.
17. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. Stat Med 1993;12: 1293–316.
18. Jordan JA. Colposcopy in the diagnosis of cervical cancer and precancer. Clin Obstet Gynecol 1985;12:67–76.
19. Powell JL. Pitfalls in cervical colposcopy. Obstet Gynecol Clin North Am 1993;20:177–88.

Address reprint requests to:
Michele Follen Mitchell, MD, MS
Department of Gynecologic Oncology
University of Texas M. D. Anderson Cancer Center
1515 Holcombe Boulevard, Box 67
Houston, TX 77030
E-mail: follen_mitchell@gyn.mda.uth.tmc.edu

Received August 29, 1997.
Received in revised form December 8, 1997.
Accepted December 30, 1997.