

Can cervicography be improved? An evaluation with arbitrated cervicography interpretations

Diana L. Schneider, DrPH,^a Louis Burke, MD,^b Thomas C. Wright, MD,^c Mark Spitzer, MD,^d Nilanjan Chatterjee, PhD,^e Sholom Wacholder, PhD,^f Rolando Herrero, MD, PhD,^g Maria C. Bratti, MD,^g Mitchell D. Greenberg, MD,ⁱ Allan Hildesheim, PhD,^j Mark E. Sherman, MD,^k Jorge Morales, MD,^l Martha L. Hutchinson, MD,^m Mario Alfaro, MD,ⁿ Attila Lörincz, PhD,^o and Mark Schiffman, MD^p

Washington, DC, Boston, Mass, New York and Manhasset, NY, Bethesda, Baltimore, and Gaithersburg, Md, Lyon, France, San Jose, Costa Rica, Philadelphia, Pa, and Providence, RI

OBJECTIVE: The purpose of this study was to estimate the optimal performance of cervicography. We compared an arbitrated cervigram classification with an arbitrated referent diagnosis of cervical neoplasia.

STUDY DESIGN: From an initial group of 8460 women, a stratified sample of cervigrams from 3645 women and histologic information from 414 women underwent arbitration. Interobserver agreement was assessed for cervicography and the referent diagnosis. Sensitivity, specificity, and predictive values were estimated for initial and arbitrated cervicography results, compared with the initial and arbitrated referent diagnoses.

RESULTS: For the detection of arbitrated high-grade lesions or cancer, arbitrated cervicography yielded an overall sensitivity of 63.9% and a specificity of 93.7%. Significantly higher sensitivity was associated with younger age and age-related visual characteristics.

CONCLUSION: Optimization of the cervigram classification improved performance over a single interpretation in this population but suggested the limits of static visual screening. (Am J Obstet Gynecol 2002;187:15-23.)

Key words: Cervicography, cervical cancer, screening, cervical neoplasia

The present research project was conducted to provide an independent evaluation of cervicography as a primary screening method for the early identification of cervical neoplasia. This evaluation of cervicography was conducted as part of a population-based study of the natural

history of cervical neoplasia in Guanacaste, Costa Rica, sponsored by the National Cancer Institute. The Guanacaste site was selected because of its consistently high age-adjusted rates of cervical cancer, despite existing Papanicolaou smear screening services.¹ Visual,² microscopic,^{3,4} and molecular⁵ screening techniques are under study.

The goal of our screening efforts was to detect high-grade cervical intraepithelial lesions (CIN2, CIN3) and cancer. In previous work,² we reported initial findings from the enrollment phase of this study that indicated that cervicography was less sensitive, and only marginally more specific, than conventional cytologic testing for the detection of high-grade lesions or cancer. However, cervicography in this study was easy to perform, with few technically defective results. It was judged potentially important, especially if sensitivity for detecting high-grade lesions could be increased without substantially reducing specificity. In an attempt to achieve a cervicography classification that approaches the optimal achievable result, we assessed whether the performance of cervicography can be improved by additional evaluation of cervigrams.

From the Division of Immigration Health Services, US Public Health Service,^a Beth Israel Medical Center,^b Columbia University,^c North Shore Hospital,^d the Statistical Research and Applications Section^e and the Biostatistics Branch,^f National Cancer Institute, Proyecto Epidemiológico Guanacaste FUCODOCSA,^g the Ministerio de Salud,^h Omnia, Inc,ⁱ the Division of Cancer Epidemiology and Genetics, National Cancer Institute,^j Johns Hopkins University,^k the Departamento de Ginecología, Caja Costarricense de Seguro Social,^l Women and Infants' Hospital of Rhode Island,^m the Departamento de Patología, Caja Costarricense de Seguro Social,ⁿ Digene Corporation,^o and the Environmental Epidemiology Branch, National Cancer Institute.^p

Supported by the National Cancer Institute, contracts No. N01-CP-21081 and NO 1-CP-31061.

Received for publication April 5, 2001; revised October 15, 2001; accepted January 9, 2002.

Reprint requests: Diana L. Schneider, DrPH, US Public Health Service, Division of Immigration Health Services, 801 I St, NW, Suite 910, Washington, DC 20536.

*© 2002, Mosby, Inc. All rights reserved.
0002-9378/2002 \$35.00 6/1/122848
doi:10.1067/mob.2002.122848*

Table I. Cervigram classification*

<i>Classification</i>	<i>Explanation</i>
<i>Not referred for colposcopy</i>	
Negative	No definite lesion is visible
Atypical 1 (A1)	A lesion inside the transformation zone is visible; based on the lesion's site and morphologic condition, the lesion is presently considered to be of doubtful significance.
Atypical 2 (A2)	A lesion outside the transformation zone is visible; based on the lesion's site and morphologic condition, the lesion is presently considered to be of doubtful significance
Technically defective	The cervigram slide is not adequate for evaluation.
<i>Referred for colposcopy</i>	
Positive (all categories below)	A lesion is visible, and colposcopy is recommended because of the lesion's site and morphologic condition or because no definite lesion is visible, but the appearance warrants colposcopy to exclude significant disease.
Positive 0 (P0)	Probably normal variant; appearance warrants colposcopy to exclude significant disease.
Positive 1A (P1A)	A lesion extends into the canal, the visible portion of which is presently considered to be of doubtful significance.
Positive 1B (P1B)	The appearance is compatible with a low-grade lesion.
Positive 2 (P2)	The appearance is compatible with a high-grade lesion.
Positive 3 (P3)	The appearance is compatible with cancer.

*As of January 1, 1995, National Testing Laboratories Worldwide revised the atypical category. Before January 1, 1995, atypical 1 referred to trivial lesions outside the transformation zone and atypical 2 referred to trivial lesions inside the transformation zone. The current terminology is applied to all cervigram classifications in this article.

Therefore, we submitted a subsample (43%) of cervigrams that were taken during enrollment for additional evaluation by an independent evaluator and subsequent arbitration by a third cervigram evaluator.

Similarly, a subsample of histologic material that was collected during the enrollment study underwent additional interpretation by independent pathologists to ascertain whether errors in original histopathologic interpretation might have accounted for low cervicography sensitivity.

Another independent histologic reading that was performed in Costa Rica for clinical purposes was used to arbitrate discrepancies between the initial referent diagnosis and the present review. To estimate "optimal" performance, we compared the arbitrated cervicography result with the arbitrated referent diagnosis.

Material and methods

Enrollment study methods. The follow-up study design, subject selection, participation rates, collection of clinical specimens, assignment of enrollment screening test results, colposcopic referral, and initial referent diagnosis are described in greater detail elsewhere.^{1,2} The protocol for this study was approved by the Institutional Review Boards of Costa Rica and the National Cancer Institute.

At enrollment into the follow-up study, cervigrams were obtained for 9062 women, which corresponded to 98.8% of women who underwent the pelvic examination. The 602 women who had undergone hysterectomy (6.6%) were subsequently excluded from the analyses because of their lack of a cervix and subsequent low risk, which left 8460 participants who were available for the enrollment phase of the cervicography evaluation study.

Two types of cytologic preparations were made for each participant, including a conventional Papanicolaou smear and a ThinPrep (Cytoc, Boxborough, Mass). After the smear was made, the Cervex brush (National Testing Laboratories, Fenton, Mo) was rinsed in 20 mL of PreservCyt (Cytoc). Vials that contained the PreservCyt solution were sent to the United States where ThinPrep slides were made. Human papillomavirus DNA results are reported elsewhere.^{5,6}

The cervix was then rinsed with 5% acetic acid, and 2 photographic images of the cervix (Cervigrams [National Testing Laboratories Worldwide]) were taken with a Cerviscope (National Testing Laboratories Worldwide [NTL], Fenton, Mo). The undeveloped film was sent to the United States to be developed, processed, and evaluated. Cervigrams that were taken during the enrollment pelvic examination were interpreted by a certified evaluator (M. D. G.) and classified according to the diagnostic criteria approved by NTL (Table I).

Three methods were used for the cytologic diagnosis: conventional Papanicolaou smear (M. A.); PapNet⁴ (Neuromedical Systems, Inc, Suffern, NY [now TriPath, Elon, NC]), which uses the same slide as the Papanicolaou smear (M. E. S.); and ThinPrep³ (M. L. H.). The Papanicolaou smear, ThinPrep, and PapNet results were classified according to the Bethesda system as negative, atypical squamous cells of undetermined significance, low-grade squamous intraepithelial lesion, high-grade squamous intraepithelial lesion, or carcinoma.⁷ Glandular lesions were rare and were classified with the closest squamous diagnosis (eg, adenocarcinoma was combined with squamous carcinoma).

Participants were referred for colposcopy if (1) physical examination was suspicious for cervical cancer, (2) there

Table II. Selection categories for cervicography review*: sampling fractions, number selected, and multipliers were used to reconstitute the original study sample

<i>Criterion</i>	<i>Percent selected (%)</i>	<i>No.</i>	<i>Multiplier†</i>
High-grade lesion or cancer at enrollment	100	136	1.0
At least 1 abnormal screening test at enrollment	100	1610	1.0
Tested positive for HPV (by hybrid capture I) at enrollment	100	298	1.0
Five or more lifetime sexual partners	100	388	1.0
Selected as control subjects for follow-up study†	100	513	1.0
Did not meet criteria	12.7	700	7.879
Total		3645	
Selection for histologic review			
Histologic CIN1 or more severe on initial review in the United States or Costa Rica	100	294	1.0
Had biopsy (less severe than CIN1 in both the United States and Costa Rica), randomly selected	34.5	120	2.903
Total		414	

*Selection for cervicography was established hierarchically; selection criteria are mutually exclusive. †Five hundred thirteen low-risk women were randomly selected for annual follow-up as a part of an ongoing prospective study. Cervigrams for all of these women were reviewed for the present analyses.

‡The multiplier is the inverse of the sampling fraction used in reconstituting the full population estimates (see Data analysis).

was an abnormal cytologic result by any of the 3 methods (atypical squamous cells of undetermined significance or more severe), or (3) there was a positive cervigram (Table I). Colposcopy was performed by a single gynecologist (J. M.) who took a biopsy specimen from the colposcopically most abnormal area, if any were visible, and took endocervical curettages as appropriate. During the colposcopy examinations, digital images (Denvu Ltd, Tucson, Ariz) of the cervix were taken for each woman that corresponded to (1) low magnification before the application of 5% acetic acid (which provides an acetowhitening effect that highlights lesions), (2) low magnification after the application of 5% acetic acid, (3) high magnification after application of 5% acetic acid, and (4) an orienting image of the biopsy site, if applicable.

As a quality-control measure, a random sample of 2% of all women was referred for colposcopy to validate the screening protocol. None of 144 women with negative screening results for all screening tests had a referent diagnosis of CIN 2 or worse.

Histologic material that included punch biopsy specimens, subsequent excisional biopsy specimens and curettages, and hysterectomy specimens was sent to the United States for review and assignment of the referent diagnosis (M. E. S.). Additionally, histologic material was diagnosed in Costa Rica for clinical purposes. Participants with a histologically confirmed high-grade squamous intraepithelial lesion or cancer or with a diagnosis of high-grade squamous intraepithelial lesion by at least 2 cytologic methods were referred for treatment through the Costa Rican Social Security system.¹ The enrollment referent diagnoses were made on the basis of histologic, cytologic, and cervicography results, with a specific diagnostic algorithm.

Additional reviews of clinical materials for this analysis

Subsequent cervigram reviews and the assignment of revised results. Cervigrams from the enrollment phase of

the study were reviewed by 1 or 2 certified evaluators (different from the evaluators at enrollment, according to the algorithm that will be discussed later in the article) to assess possible errors in the initial interpretation. The cervigrams were selected for review on the basis of previous screening or diagnostic outcomes and/or risk factors. The targets of review were positive cervigrams and high-grade disease outcomes. The sample composition and corresponding sampling fractions for the cervigram review are shown in Table II.

Evaluators were masked from knowing the previous cervigram, cytologic, histologic, or human papillomavirus results or even the composition of the sample. However, they were aware of the general results from the enrollment study. Cervigrams that corresponded to 3645 women were selected for review.

During the review, cervigrams were again classified into the categories shown in Table I. We compared the cervigram results between the initial evaluator at enrollment and the second evaluator (M. S.). Cervigrams that corresponded to women whose classification differed (based on classification as negative, atypical, positive 0, positive 1, positive 2, positive 3, or technically defective) between the 2 evaluators (n = 820 cervigrams; 22.4%) were sent, along with a 10% sample of cervigrams for women with concordant results (n = 282 cervigrams), to a third evaluator (L. B.) for arbitration. The subsample of women with concordant results was included to mask the sample composition. A revised cervigram result was assigned based on the agreement of 2 of the 3 evaluators. For the analyses presented here, a positive cervigram result includes all categories of positive cervigrams (ie, positive 0, positive 1, positive 2, and positive 3 versus negative or atypical).

Additional information was recorded about the cervigram in an attempt to explain discordant results and to stratify cervicography performance estimates. Visual

Table III. Observer agreement on cervigram and histologic classification

Arbitrated cervigram result	Cervigram result assigned by initial reviewer			κ statistic
	Positive	Negative, atypical, or technically defective	Total	
Positive	388	99	487	0.8
Negative, atypical, or technically defective	94	3056	3150	
Total	482	3155	3637	

High-grade lesions include cervical intraepithelial neoplasia grade 2 and cervical intraepithelial neoplasia grade 3; low-grade lesions include cervical intraepithelial neoplasia grade 1 and koilocytotic atypia.

characteristics that were recorded included (1) whether the lesion was seen in its entirety; (2) whether columnar epithelium was visible on the ectocervix; (3) whether metaplasia was visible on the ectocervix; (4) whether the cervigram showed a congenital transformation zone; (5) the presence and quality of an acetic acid effect; (6) whether the transformation zone was partially obscured by blood, mucus, the position of the cervix, hair, vaginal wall, or speculum; and (7) the presence and degree of inflammation.

Digital colposcopic image review. We assessed the possibility of error in the colposcopy examination during the enrollment study through a review (L. B.) of the digital images taken during the examination (Denvu Ltd). All available images that were taken during the initial colposcopy examination were included in the review ($n = 1983$ women; 96.4% of all colposcopy examinations). Images were evaluated according to the reviewer's agreement or disagreement with the decision to take a biopsy specimen and, if the specimen had been taken, on the reviewer's agreement or disagreement with the biopsy placement.

Subsequent histologic review and assignment of the arbitrated referent diagnosis. We assessed the possibility of error on the referent diagnosis during the enrollment study. An independent pathologist in the United States (T. C. W.) reviewed a sample of histologic slides from the enrollment study, and a diagnostic classification was recorded. Slides that were selected for the review sample ($n = 414$ cases) included those women for whom the histologic result during enrollment, as interpreted either in the United States or in Costa Rica, corresponded to cervical intraepithelial neoplasia 1 (CIN 1) or more severe, plus a random sample of slides that corresponded to the remaining women with available histologic slides whose most severe histologic result during enrollment was normal or equivocal. The sample composition and corresponding sampling fractions for the histologic review are shown in Table II.

The referent diagnosis was grouped as high-grade squamous intraepithelial lesions (corresponding to CIN2 and CIN3) or cancer versus low-grade squamous intraepi-

thelial lesions (corresponding to CIN 1 and koilocytotic atypia) or less severe (normal, atypical, or equivocal). In the enrollment study,¹ we also described 8 women with definite cytologic diagnoses of high-grade squamous intraepithelial lesions without histologic confirmation. For this analysis, we categorized these 8 women as having low-grade lesions or less severe lesions to provide an assessment of cervicography compared only with histologically confirmed high-grade squamous intraepithelial lesions or cancer.

An arbitrated referent diagnosis was provided according to the following protocol: The arbitrated referent diagnosis was based primarily on an agreement of 2 histologic reviews. After the histologic arbitration process, the referent diagnoses were classified as (1) cancer, (2) histologically confirmed high-grade squamous intraepithelial lesions, (3) histologically confirmed low-grade squamous intraepithelial lesions, or (4) normal, atypical, or equivocal. In the cases for which the second histologic diagnosis was the same as the initial diagnosis, the referent diagnosis for that participant was not reassigned. In the cases for which the diagnoses at enrollment and the subsequent review differed, the histologic diagnoses that were provided by the pathologists in Costa Rica during enrollment (which guided clinical treatment) were used to arbitrate.

Data analysis. Women who were selected for inclusion in the cervigram review for whom the pair of cervigrams were not available ($n = 6$ women) or both cervigrams were uninterpretable by a reviewer were excluded from all analyses ($n = 2$ women). These exclusions included 2 women with a referent diagnosis of high-grade squamous intraepithelial lesions at enrollment. Additionally, 14 women were excluded from relevant analyses because of unavailable histologic results, lack of identification, or noninclusion in the cervicography review subsample. After all exclusions, 3637 arbitrated cervigram results and 400 arbitrated referent diagnostic results were available for analysis.

Sensitivity, specificity, percent of women referred for colposcopy, and positive and negative predictive values were initially calculated with standard contingency table methods,⁸ which compared the enrollment screening test

<i>Arbitrated referent diagnosis</i>	<i>Histologic result assigned by initial reviewer</i>			<i>κ statistic</i>
	<i>High-grade, or cancer</i>	<i>Negative, atypical, equivocal, low-grade, or other</i>	<i>Total</i>	
Arbitrated referent diagnosis				
High-grade or cancer	112	3	115	0.9
Negative, atypical, equivocal, low-grade, other	16	269	285	
Total	128	272	400	

results with the initial referent diagnosis as the gold standard. A detailed report of the findings from the enrollment study is available elsewhere.² Small differences in total numbers may be found between the present and previous analyses because of exclusions that were specific to the analysis of the arbitrated results.

For women whose cervigrams and histologic slides were included in the review samples, we assessed interobserver agreement between the results by the initial reviewer at enrollment and by the second reviewer and between the initial and arbitrated classifications, using the kappa statistic. We interpreted the kappa statistics with the scale described by Altman.⁹ Briefly, this scale classifies agreement beyond that expected by chance alone as “very good” if κ is 0.81 to 1.00, as “good” if κ is 0.61 to 0.80, as “moderate” if κ is 0.41 to 0.60, as “fair” if κ is 0.21 to 0.40, and as “poor” if κ is <0.20 .

We estimated sensitivity, specificity, and positive and negative predictive values for cervicography after we incorporated arbitrated cervicography and histologic results with the maximum likelihood estimate of the joint distribution to account for sampling (Appendix). In a separate approach, we reconstituted the original sample of women (from the enrollment phase of the study) according to the sampling fractions by which women were selected into the cervicography reviews (Table II). Reconstitution was achieved by multiplying the contingency table frequencies from the review sample by the inverse of the sampling fraction that corresponded to each category of selection into the review phase. In other words, these analyses were stratified by the categories that had been established for selection into the cervigram reviews. Each stratum was reconstituted, and stratum-specific results were combined before the analyses were performed. For the analyses that included arbitrated cervigram and arbitrated referent diagnoses, the results yielded by reconstituting the sample were very close to those yielded by the maximum likelihood method. The overall results that are shown are those results that were yielded with the use of the maximum likelihood estimation method. We stratified sensitivity and specificity by possible predictors of

error, including age, menopause status, visual characteristics of the cervigram, reviewer agreement on the decision to take a biopsy specimen, and agreement on biopsy specimen placement. The stratified estimates described later in the article were achieved by the use of the reconstitution method. We tested for associations between these characteristics and cervigram results for women with an arbitrated referent diagnosis of high-grade lesions or cancer with the Fisher exact test.¹⁰

Results

Observer agreement on cervigram classification. The cervigram classifications that were assigned by the initial evaluator versus the second evaluator and by the initial evaluator versus the arbitrated result were compared for the 3637 women who were included in the cervigram review and for which results were available. A comparison of dichotomous results that were assigned by the initial versus second evaluator yielded a kappa statistic of 0.5, which indicated only moderate agreement beyond that expected by chance. In contrast, the cervigram classification that was assigned by the initial evaluator compared with the arbitrated classification yielded a κ statistic of 0.8, which indicated good agreement beyond that expected by chance alone (Table III).

Table III also shows that agreement between the initial and arbitrated referent histologic diagnosis is slightly better than that observed for cervicography. There were 19 discrepant histologic results, 16 of which were downgraded from high-grade lesions or cancer initially to low-grade lesions or less severe on arbitration.

Cervicography screening compared with the referent diagnosis. The 2 referral categories of cervigram classification (positive [ie, referred for colposcopy] versus normal, atypical, or technically defective [ie, not referred]) were used to determine percent referred, sensitivity, specificity, and predictive values of cervicography.

Table IV presents the sensitivity, specificity, positive predictive value, and negative predictive value for the initial and arbitrated cervigram results compared with the initial and arbitrated referent diagnoses. These analyses demonstrate that, when the cervicography and referent diagnosis

tic results were optimized, the sensitivity of cervicography improved by 12% compared with the sensitivity yielded by a single interpretation of each test. Specifically, arbitration of histologic condition further increased the sensitivity of arbitrated cervicography to 63.9%. A deceptively slight reduction in specificity was noted with the optimized results. However, this 1.3% reduction in specificity would increase the overall proportion of women referred for colposcopy from 5.7% to 7.1%, which would result in almost 25% more referrals in relative terms. Of the 11 cases of invasive cancer, 10 cases (90.9%) were identified by the arbitrated cervicography process. It is noteworthy that the 1 case of invasive cancer that was not detected by the arbitrated cervigram result was initially correctly identified by the evaluator at enrollment.

The effects of various characteristics on cervigram results. To further evaluate the performance of cervicography, we assessed the sensitivity and specificity of the arbitrated cervigram result that had been stratified by various characteristics of the women and their cervigrams. We used stratification and not maximum likelihood estimation for this part of the analysis, which slightly affected the estimates. Stratification was performed in an attempt to explain the reasons for the relatively limited detection of high-grade lesions that were observed in the overall results. Sensitivity (based on the arbitrated referent diagnosis) was significantly higher among the 6478 women who were younger than age 50 years (66.3%; Fisher exact test, $P = .02$) compared with the 1969 women aged ≥ 50 years (36.8%) and, correspondingly, among the 6280 women who were premenopausal (67.0%; Fisher exact test, $P = .02$) compared with the 2167 women who were postmenopausal (39.1%). Of the age-related visual characteristics on the cervigram that were assessed (eg, atrophy, metaplasia, and acetic acid effect), only an increasing quality of the acetic acid effect was statistically significantly associated with higher sensitivity (Fisher exact test, $P < .001$).

The presence of the following age-related characteristics resulted in significantly lower specificity of cervicography (all $P < .001$, Fisher exact test): premenopausal status, entirety of the lesion not visible, metaplasia visible, altered columnar epithelium visible, presence of a congenital transformation zone, absence of cervicovaginal atrophy, the presence or better quality of an acetic acid effect, and a lack of appearance of friability. Statistical significance of many small differences in specificity may be explained by the higher statistical power of these analyses, because of the larger numbers of women without serious neoplasia. For example, specificity was statistically significantly reduced in women whose cervigrams showed a congenital transformation zone, although few women had this characteristic ($n = 39$ women).

Digital colposcopic images were available for 1983 women (96.4% of all women who underwent colposcopy). The 2 key variables that were assessed included

agreement on the decision to perform a biopsy and agreement on biopsy placement within 5 mm (if a biopsy specimen was taken). Of these 1983 women, 320 women (16.1%) had a biopsy specimen taken, and 1615 women (81.4%) did not (the images for an additional 45 women [2.3%] were insufficient for assessment; the response for 3 women [0.2%] was missing). The reviewer (L. B.) agreed with the decision to perform a biopsy for 224 of the 320 women (70.0%) who had a biopsy report; the reviewer agreed with the decision not to take a biopsy specimen for 1233 of the 1615 women (76.3%) who did not have a biopsy specimen taken. The image reviewer agreed with biopsy placement for 170 of the 224 women (75.9%) for whom a biopsy specimen was taken, and there was agreement on the decision to perform a biopsy.

Estimates of sensitivity and specificity were stratified by the agreement on the decision to perform a biopsy and the biopsy site. We found that the sensitivity of cervicography was nonsignificantly higher (73.5%) in women for whom there was non agreement on the decision to perform a biopsy, compared with women for whom there was agreement on the decision to perform a biopsy (60.0%; $P = .3$). Of the 380 women for whom the digital colposcopic image review revealed lack of agreement on the decision to perform a biopsy, 294 women (77.4%) had not had a biopsy specimen taken, and the reviewer indicated that a biopsy specimen should have been taken (ie, the original colposcopic examination might have failed to detect a high-grade lesion); 86 women (22.6%) had a biopsy specimen taken, and the reviewer indicated that a biopsy specimen need not have been taken. Among women who had a biopsy specimen taken and for whom the image was adequate for assessment ($n = 218$ women), no significant difference in sensitivity was observed between women for whom the digital colposcopic image review indicated agreement on biopsy placement (71.4%) and women for whom there was no agreement on biopsy placement (50.0%; $P = .6$). In the subgroup of women for whom the referent diagnosis was arbitrated, the decision to perform a biopsy was corroborated, and the choice of biopsy site was confirmed ($n = 90$ women), cervicography yielded a sensitivity of 71.4% and a specificity of 51.6%.

Comment

In our initial evaluation of cervicography as a primary screening test for cervical neoplasia,² we established that cervicography had imperfect sensitivity. During the enrollment study, 5.7% of the 8460 women were referred for colposcopic examination because of a positive cervigram. Cervicography resulted in the detection of all 11 cases of invasive cervical cancer and 49.3% of high-grade squamous intraepithelial lesions and cancer combined (with the initial referent diagnosis as the gold standard).

Table IV. Initial and arbitrated cervicography results compared with the initial and arbitrated referent diagnosis

	<i>High-grade* or cancer</i>	<i>Normal, equivocal, low-grade†</i>	<i>Total (n)</i>	<i>Sensitivity (%)</i>	<i>Specificity (%)</i>	<i>Predictive value (%)</i>	
						<i>Positive</i>	<i>Negative</i>
A. Initial cervigram result compared with the initial referent diagnosis‡							
Initial cervigram result							
Positive	67	417	484				
Negative, atypical or technically defective	61	7915	7976	52.3	95.0	13.8	99.2
Total	128	8332	8460				
B. Arbitrated cervigram result compared with the initial referent diagnosis§							
Arbitrated cervigram result							
Positive	75	527	602				
Negative, atypical or technically defective	53	7805	7858	58.6	93.7	12.5	99.3
Total	128	8332	8460				
C. Initial cervigram result compared with the arbitrated referent diagnosis							
Arbitrated referent diagnosis							
Initial cervigram result							
Positive	51	433	484				
Negative, atypical, or technically defective	71	7905	7976	41.8	94.8	10.5	99.1
Total	122	8338	8460				
D. Arbitrated cervigram result compared with the arbitrated referent diagnosis							
Arbitrated cervigram result							
Positive	78	524	602				
Negative, atypical, or technically defective	44	7814	7858	63.9	93.7	13.0	99.4
Total	122	8338	8460				

*High-grade lesions include CIN2 and CIN3.

†Low-grade lesions include CIN1 and koilocytotic atypia.

‡Estimates were derived with the use of the maximum likelihood estimates of the joint distribution.

§Differences in column totals are due to exclusions in the cervicography and histologic reviews because of missing or uninterpretable cervigrams or because of histologic results that were not available from all 3 sources, that could not be linked to an identification number, or for whom histologic results were reviewed but did not meet the criteria for inclusion in the cervicography review subsample.

The specificity of cervicography was 95.0%, with a positive predictive value of 13.8% and a negative predictive value of 99.1%. These results led to the design of the present study in which we assessed whether cervicography screening could be improved with additional interpretation that was followed by an arbitration process.

Cervicography suffers from imperfect reproducibility. In a comparison of the cervigram classification between the initial and second evaluators, results showed moderate agreement beyond that expected by chance alone, similar to the cytologic diagnosis. However, evaluator agreement beyond chance was good when a comparison was made of the initial and arbitrated cervicography classification, which suggests that arbitration did not greatly change the initial results. Studies of interobserver agreement of the Papanicolaou smear have also shown moderate or even poor reproducibility.¹¹⁻¹³

The arbitration process, with both cervicography and histologic findings optimized, not only yielded improved sensitivity over the initial findings that were based on single evaluations but also produced lower specificity and positive predictive value. The arbitrated cervigram classification resulted in 7.1% of women being referred for colposcopic examination, and a sensitivity of 63.9%, a specificity of 93.7%, a positive predictive value of 13.0%, and a negative predictive value of 99.4% for the detection of cervical intraepithelial neoplasia grades 2 and 3 and cancer. The value of this improvement compared with its

cost must be weighed but is beyond the scope of this analysis.

As a point of comparison, during the enrollment phase of our study, conventional cytologic screening resulted in 6.9% of women being referred for colposcopy and a sensitivity of 77.2%, a specificity of 94.2%, a positive predictive value of 17.9%, and a negative predictive value of 99.6%.² Optimization of cytologic smear procedures before the start of the study may partially explain the reason that conventional cytologic diagnoses in our study performed better than in previous studies.¹¹

We stratified sensitivity and specificity by characteristics of the woman and of her cervigram in an attempt to further understand the performance of cervicography. Characteristics of the cervigram were noted by the second and third cervigram evaluators only. Because many of these factors were not assessed during the initial enrollment study, we were unable to determine evaluator agreement on these characteristics. Cervicography is significantly more sensitive in women younger than age 50 years and in women who are premenopausal than in women aged ≥ 50 years and women who were postmenopausal, respectively. The marked reduction in sensitivity in women after menopause is associated with positional change in the transformation zone. Most cervical neoplasia occurs at the transformation zone, which moves cephalad into the endocervical canal as a woman ages. Because the cervigram evaluator visualizes the pro-

jected image of the cervix, the technique does not allow for the detection of lesions completely inside the endocervical canal. The usefulness of cervicography in women after menopause and/or in women aged ≥ 50 years is therefore very limited.

Of the visual characteristics of the cervigram that were observed, only the presence of and increasing quality of the acetic acid effect were associated with sensitivity. This association was expected because the acetic acid produces the acetowhitened highlighting of cervical abnormalities. Several of the visual characteristics that were observed were significantly associated with the specificity of cervicography. Statistical significance that was achieved from some apparently small differences in specificity may be explained by the high statistical power of these analyses, because of the very large numbers of women without serious neoplasia. The clinical importance of these influences on specificity is likely to be small. Of note, many of the apparent false positive results were for women who were classified to be human papillomavirus DNA negative.²

Limitations of this study are that not all cervigrams and histologic specimens that were evaluated during the enrollment study were re-evaluated and that not all cervigrams that were reviewed underwent arbitration. The cost of evaluating all cervigrams and histologic findings in triplicate would have been prohibitive for this study. Nonetheless, we feel that we developed reasonable estimates of optimized cervigram performance that were relative to a referent diagnostic classification.

Cervigrams in this study were reported (by the third evaluator) to show more blood than is usually seen in cervigrams. Excessive bloodiness may have contributed to high false positivity that is associated with positive 0 cervigrams because bloodiness is one of the guiding criteria for a positive 0 classification. In our study, of the 46 women with a revised cervigram result of positive 0, 17 women (37%) had cervigrams reported with the transformation zone that was partially obscured by blood. It is possible that bleeding may have been due to the specimen collection protocol that required cervicography photographs to be taken after cytologic sampling and therefore may have been affected by excessive scraping. In our protocol, cervicography was not performed before cytologic sampling because of a concern that the application of acetic acid might interfere with the cytologic results. An alternate explanation is that at least some of the bleeding was associated with high rates of cervical inflammation in this population. This explanation is partially supported by the finding that 30% of all women in this study were observed to have some degree of cervical inflammation on the basis of the appearance of their cervigrams. A study that was based on microscopic assessment has confirmed this impression.¹⁴

Digital colposcopic images were reviewed to assess agreement with the initial colposcopy result. However, it

should be noted that the quality of these images was deemed by the reviewer to be less than optimal, so the results that correspond to the digital colposcopic image review should be interpreted with caution. The evaluator of these colposcopic images (not the cervigram) noted an apparent deficiency in the application of acetic acid, and mucus was not adequately removed from many of these images, thus impairing the visualization of the cervix. Additionally, it was not always clear which part of the cervix was visible in the image. This was more common with low-grade lesions than with high-grade lesions. These observations are believed to reflect more on inadequate application of acetic acid and documentation than on the quality of the digital colposcopy technique.

A strength of this study is the large, population-based sample in which it was conducted. The large sample permitted the identification of a sufficiently large number of women with high-grade squamous intraepithelial lesions or cancer to assess the performance of cervicography for these serious lesions separately from low-grade squamous intraepithelial lesions. There was extensive training, and virtually all cervigram images were judged to be high quality. The reviews of cervigrams and histologic material allowed us to estimate optimal cervigram and referent diagnostic results for each woman and to stratify by potential reasons for misclassification.

In summary, this study suggests that the optimal sensitivity of cervicography that is based on arbitrated reviews could be improved moderately over cervicography screening with the use of single cervigram and referent diagnostic evaluations, at the expense of slight reductions in specificity and positive predictive value. Cervicography is subject to fair-to-moderate interobserver agreement, and it detects fewer high-grade cervical lesions than does the Papanicolaou smear in this mass screening setting. However, sensitivity is high for the detection of invasive cancer and is similar to that of conventional cytologic diagnoses. Cervicography alone is of limited use in women aged ≥ 50 years and in women after menopause, because the sensitivity drops markedly in these groups. In future work, we are now examining possible complementary screening methods in combination, including cervicography and cytologic diagnoses, for general screening in regions with different health resources. In parallel, in a US study population, we are examining the ancillary use of cervicography for the triage of unclear cytologic diagnoses.¹⁵

We thank the Costa Rican study team, especially Ileana Balmaceda, Lidia Ana Morera, Fernando Cárdenas, Manuel Barrantes, and Elmer Perez for their hard work and dedication and Linda Saxon, Julie Buckland, and Pei Chao of Information Management Services, Inc (Silver Spring, Md) for their invaluable computer support.

REFERENCES

1. Herrero R, Schiffman MH, Bratti C, Hildesheim A, Balmaceda I, Sherman ME, et al. Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa Rica: the Guanacaste project. *Pan Am J Public Health* 1997;1:362-75.
2. Schneider DL, Herrero R, Bratti C, Greenberg MD, Hildesheim A, Sherman ME, et al. Cervicography screening for cervical cancer among 8,460 women in a high-risk population. *Am J Obstet Gynecol* 1999;180:290-8.
3. Hutchinson M, Zahniser D, Sherman ME, Herrero R, Alfaro M, Bratti C, et al. Utility of liquid-based cytology for cervical carcinoma screening: results of a population-based study conducted in a region of Costa Rica with a high incidence of cervical carcinoma. *Cancer Cytopathol* 1999;87:48-55.
4. Sherman ME, Schiffman MH, Herrero R, Alfaro M, Kelly D, Bratti C, et al. Performance of a semi-automated Papanicolaou smear screening system: results of a population-based study conducted in Guanacaste, Costa Rica. *Cancer Cytopathol* 1998;84:273-80.
5. Schiffman M, Herrero R, Hildesheim A, Sherman ME, Bratti C, Wacholder S, et al. HPV DNA testing in cervical cancer screening: results from women in a high-risk province of Costa Rica. *JAMA* 2000;283:87-93.
6. Herrero R, Hildesheim A, Bratti C, Sherman ME, Hutchinson M, Morales J, et al. Population-based study of human papillomavirus infection and cervical neoplasia in rural Costa Rica. *J Natl Cancer Inst* 2000;92:464-74.
7. Kurman RJ, Soloman D. The Bethesda system for reporting cervical/vaginal cytologic diagnoses: definitions, criteria, and explanatory notes for terminology and specimen adequacy. New York: Springer-Verlag; 1994.
8. Fleiss JL. Statistical methods for rates and proportions. New York: John Wiley; 1981.
9. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991.
10. Daniel WW. Biostatistics: a foundation for analysis in the health sciences. New York: John Wiley; 1995.
11. Fahey MT, Irwig L, Macaskill P. Meta-analysis of Pap test accuracy. *Am J Epidemiol* 1995;141:680-9.
12. Sherman ME, Schiffman MH, Lorincz AT, Manos MM, Scott DR, Kurman RJ, et al. Toward objective quality assurance in cervical cytopathology. *Am J Clin Pathol* 1994;102:182-7.
13. Stoler MH, Schiffman M, for the ALTS group. Inter-observer reproducibility of cervical cytologic and histologic diagnoses: realistic estimates from the ASCUS-LSIL triage study (ALTS). *JAMA* 2001;285:1500-5.
14. Castle P, Hillier S, Rabe LK, Hildesheim A, Herrero R, Bratti MC, et al. An association of cervical inflammation with high-grade cervical neoplasia in women infected with oncogenic HPV. *Cancer Epidemiol Biomarkers Prev* 2001;10:1021-7.
15. Ferris DG, Schiffman M, Litaker MS, for the ALTS group. The sensitivity of cervicography for triage of women with ASCUS and LSIL based on ASCUS LSIL triage study enrollment data. *Am J Obstet Gynecol* 2001;185:939-45.

Appendix

Maximum-likelihood estimation. We used maximum-likelihood estimation to obtain estimates of sensitivity, specificity, positive predictive value, and negative predictive value that would apply to the entire study population

from our stratified sampling plan. Some notation will be useful to describe the statistical method. Let x_0 , x_1 , y_0 , and y_1 denote the results from the initial cervicography, arbitrated cervicography, initial referent diagnosis, and arbitrated referent diagnosis, respectively, each with possible values 0 (absent) or 1 (present). Note that, the specificity and sensitivity of cervicography compared with the referent diagnosis can be computed as a function of the joint frequency distribution of the 2 tests. Because the joint distribution of any pair of tests can be easily obtained once we have an estimate for the joint distribution of all the 4 tests together, let us consider the estimation of the latter. Let $P(x_0, x_1, y_0, y_1)$ denote the joint distribution of the 4 tests together. For N individuals, $i = 1, \dots, N$, we have (x_0, y_0) for everybody, but x_1 and/or y_1 only for selected individuals. Let C_{00} , C_{01} , C_{10} , and C_{11} denote the set of individuals for whom none, only y_1 , only x_1 , and both of the tests are available, respectively. The probabilities or the likelihood for the observed test results for the individuals in these 4 sets are given by $P(x_0, y_0)$, $P(x_0, y_0, y_1)$, $P(x_0, y_0, x_1)$, and $P(x_0, y_0, x_1, y_1)$, where the joint distribution of 2 and 3 tests are obtained by marginalization of the distribution of all the 4 tests over the unobserved test(s). Thus, the likelihood of the whole data can be written as

$$L = \prod_{C_{00}} P(x_0, y_0) \prod_{C_{01}} P(x_0, y_0, y_1) \prod_{C_{10}} P(x_0, y_0, x_1) \prod_{C_{11}} P(x_0, y_0, x_1, y_1). \quad (1)$$

This likelihood can be maximized with respect to the $15(2^4-1)$ parameters that define the joint distribution $P(x_0, y_0, x_1, y_1)$. However, not all 15 parameters are estimable from this data because some of the 16 possible cells are empty in this data. To overcome this problem, we considered a restricted maximum likelihood estimator. First, observe that the basic probability rule implies

$$P(x_0, y_0, x_1, y_1) = P(x_1)P(x_0 | x_1)P(y_0, y_1 | x_0, x_1) \quad (2)$$

Now, if we assume that x_1 , the arbitrated cervicography, is a better test than x_0 , the initial cervicography, in the sense that the given x_1, x_0 does not have any additional value to predict y_0 and y_1 , we will have

$$P(y_0, y_1 | x_0, x_1) = P(y_0, y_1 | x_1) \quad (3)$$

Thus, in this case $P(x_0, y_0, x_1, y_1)$ can be determined by $1 + 2 + 6 = 9$ parameters, all of which can be estimated by maximization of the likelihood of the data. The estimate of these 9 parameters then can be combined with the use of equations 2 and 3 to produce an estimate of $P(x_0, y_0, x_1, y_1)$.