

Interobserver Reproducibility of Cervical Cytologic and Histologic Interpretations

Realistic Estimates From the ASCUS-LSIL Triage Study

Mark H. Stoler, MD

Mark Schiffman, MD, MPH

for the Atypical Squamous Cells of Undetermined Significance–Low-grade Squamous Intraepithelial Lesion Triage Study (ALTS) Group

THE INTERPRETIVE REPRODUCIBILITY of cervical cytology and histopathology is critical to cervical cancer prevention programs. There is a societal presumption of high reproducibility of cytologic screening. In the medical community, histopathologic interpretations are generally considered the reference standard upon which treatment of cervical disease is based. No test or interpretation is perfect, and both society and the medical profession may have excessively high expectations. In fact, realistic standards of interpathologist agreement for cytology and histology have not been well defined by rigorous studies of large series of specimens.

Cytology screening interpretations define which women require focused clinical attention. Organized screening programs based on periodic conventional Papanicolaou (Pap) smears are successful in greatly reducing cervical cancer deaths.¹ In recent years, however, cervical cytologic screening has come under attack because of a growing awareness of the test's imperfections, including irreproducibility and false negativ-

See also p 1506.

Context Despite a critical presumption of reliability, standards of interpathologist agreement have not been well defined for interpretation of cervical pathology specimens.

Objective To determine the reproducibility of cytologic, colposcopic histologic, and loop electrosurgical excision procedure (LEEP) histologic cervical specimen interpretations among multiple well-trained observers.

Design and Setting The Atypical Squamous Cells of Undetermined Significance–Low-grade Squamous Intraepithelial Lesion (ASCUS-LSIL) Triage Study (ALTS), an ongoing US multicenter clinical trial.

Subjects From women enrolled in ALTS during 1996-1998, 4948 monolayer cytologic slides, 2237 colposcopic biopsies, and 535 LEEP specimens were interpreted by 7 clinical center and 4 Pathology Quality Control Group (QC) pathologists.

Main Outcome Measures κ Values calculated for comparison of the original clinical center interpretation and the first QC reviewer's masked interpretation of specimens.

Results For all 3 specimen types, the clinical center pathologists rendered significantly more severe interpretations than did reviewing QC pathologists. The reproducibility of monolayer cytologic interpretations was moderate ($\kappa=0.46$; 95% confidence interval [CI], 0.44-0.48) and equivalent to the reproducibility of punch biopsy histopathologic interpretations ($\kappa=0.46$; 95% CI, 0.43-0.49) and LEEP histopathologic interpretations ($\kappa=0.49$; 95% CI, 0.44-0.55). The lack of reproducibility of histopathology was most evident for less severe interpretations.

Conclusions Interpretive variability is substantial for all types of cervical specimens. Histopathology of cervical biopsies is not more reproducible than monolayer cytology, and even the interpretation of LEEP results is variable. Given the degree of irreproducibility that exists among well-trained pathologists, realistic performance expectations should guide use of their interpretations.

JAMA. 2001;285:1500-1505

www.jama.com

ity.²⁻¹² Problems with specimen collection and preparation may be partly ameliorated with monolayer preparations.^{1,13} Ultimately, cervical cytologic screening is entirely predicated on the combined judgment of the cytotechnologist and pathologist. Although clinicians

vary in their management of women with abnormal cytology, different diagnostic interpretations of any given cytologic specimen may lead to radically different patient management.

Clinicians often evaluate abnormal cytologic findings using colposcopy,

Author Affiliations: University of Virginia Health System, Charlottesville (Dr Stoler); and National Cancer Institute, Bethesda, Md (Dr Schiffman).

Members of the ALTS Group are listed at the end of this article.

Corresponding Author and Reprints: Mark H. Stoler,

MD, University of Virginia Health System, Division of Surgical Pathology and Cytopathology, Box 800214, Charlottesville, VA 22908 (e-mail: mhs2e@virginia.edu).

Toward Optimal Laboratory Use Section Editor: David H. Mark, MD, MPH, Contributing Editor.

with guided biopsies of visually abnormal areas. Colposcopy itself is not well standardized^{14,15} and the reproducibility of biopsy interpretation might actually be as variable and problematic as cytologic interpretation.^{6,16-27} Previous studies on the reproducibility of cervical preneoplasia interpretation have been limited in size and for the most part statistically inadequate.

Many clinicians now treat significant intraepithelial neoplasia, either proven or suspected, using the loop electrosurgical excision procedure (LEEP) to remove the cervical transformation zone. This procedure produces a large histology specimen that is processed similarly to a cone biopsy, oriented as a clock face in 12 sections. The resultant pathology report further defines the grade of neoplasia and guides the patient's subsequent management. Despite the widespread use of LEEP for the treatment of substantial cervical neoplasia, the reproducibility of LEEP histopathology has not been rigorously evaluated.

We evaluated the reproducibility of cytology, biopsy histopathology, and LEEP histopathology among multiple, well-trained observers in the context of an ongoing multicenter clinical trial.

METHODS

Population

The Atypical Squamous Cells of Undetermined Significance–Low-grade Squamous Intraepithelial Lesion (ASCUS-LSIL) Triage Study (ALTS) is a multicenter randomized clinical trial with 3 study arms designed to evaluate the management of mildly abnormal cytology findings by 3 alternative methods: immediate colposcopy, conservative cytologic follow-up, or triage by human papillomavirus (HPV) DNA testing.²⁸ At enrollment into ALTS during 1996-1998, women referred for ASCUS or LSIL conventional Pap smears had a repeat cytologic interpretation on monolayer cytology (Thin-Prep, Cytoc, Boxborough, Mass). Women triaged to colposcopy as required by the study protocol under-

went biopsy if lesions were visible upon application of acetic acid. Histologically confirmed cervical intraepithelial neoplasia (CIN) grades 2 to 3 was treated by LEEP. The few cases of prevalent, invasive carcinoma were treated more extensively as appropriate. A full description of the study is available elsewhere.²⁸

During enrollment, the ALTS clinical centers interpreted 4948 monolayer cytology slides, 2237 biopsies (taking only the most severe result for each woman, as described below), and 535 LEEP specimens that were independently reviewed by the Pathology Quality Control Group (QC). There were 1 to 2 staff pathologists per clinical center (7 in all), who worked independently. No conferences were held regarding cases. The initial QC review was randomly assigned to 1 of the 4 QC pathologists. The QC review was masked to the clinical center interpretation and all other test results. The present analysis is based on the comparison of clinical center to first QC interpretations. When the first QC reviewer disagreed with the original clinical center interpretation, additional reviews were performed.²⁸ While the QC algorithms and panel interpretations were used for ALTS to define disease end points, the patients were managed by the original clinical center interpretations unless CIN3 or cancer was suspected, in which case the final QC opinion was unmasked.

For cytologic specimens, cytotechnologists' screening marks were not removed during rescreening at the QC center at Johns Hopkins University. Quality control histology reviews were performed on all the original slides interpreted at the clinical centers. No recuts or substitute slides were used. Interpretations were coded using the Bethesda System squamous intraepithelial lesion (SIL) categories. Histologic and LEEP interpretations were categorized by severity of overall interpretation for a case rather than by individual block, analogous to actual clinical management. Thus, no woman contributed more than 1 interpreta-

tion to the data tables for a given specimen type. Analyses were repeated to look for trends in subgroups. These included dividing the data by each individual QC pathologist, dividing the data by each of 4 clinical centers, and analyzing the data over time to see if interpretative reproducibility varied over the period of enrollment. Finally, the results of HPV testing were briefly considered in association with the cytology and histology interpretations.²⁸

Statistical Analysis

Reactive, reparative, and inflammatory changes were grouped as negative for this analysis. The very few invasive cancer interpretations were included in the high-grade intraepithelial group. For histology, the results were combined into cytology-like (Bethesda System SIL) groupings although the CIN terminology was retained (eg, CIN2 or 3 is analogous to high-grade squamous intraepithelial lesion [HSIL]), to permit comparisons across equal-sized data tables with cytology.

κ Values were calculated to test for reproducibility while taking chance agreement into account.^{29,30} We composed 4×4 diagnostic tables, as well as more condensed 2×2 diagnostic tables, using each possible binary cutpoint (ie, negative vs \geq ASCUS, \leq ASCUS vs \geq LSIL, and \leq LSIL vs \geq HSIL). Both unweighted and weighted κ values were considered. Weighted values, with weights inversely proportional to the number of categories of distance between 2 ratings, are sensitive to severe disagreements as opposed to 1-category disagreements. Specifically, the weights were 1.00 for data cells on the diagonal (ie, exact agreement), 0.67 for cells adjacent to the diagonal, 0.33 for cells 2 units from the diagonal, and 0 for cells 3 units from the diagonal. Weighted κ values are higher than unweighted values when disagreements are common but tend to be close to the diagonal. As a rough guide, a κ value of less than 0 indicates poor agreement, 0 to 0.2 represents slight agreement, 0.2 to 0.4 is fair agreement, 0.4 to 0.6 indi-

Table 1. Interpretations: Original vs First Quality Control Group Reviewer*

Original Interpretation	First Quality Control Group Reviewer Interpretation				Total
	Negative	ASCUS	LSIL	≥HSIL	
Monolayer Cytology					
Negative					
Frequency	1325	322	52	8	1707
%	26.78	6.51	1.05	0.16	34.50
Row %	77.62	18.86	3.05	0.47	
Column %	67.46	23.94	3.93	2.52	
ASCUS					
Frequency	568	633	245	27	1473
%	11.48	12.79	4.95	0.55	29.77
Row %	38.56	42.97	16.63	1.83	
Column %	28.92	47.06	18.53	8.52	
LSIL					
Frequency	57	292	908	78	1335
%	1.15	5.90	18.35	1.58	26.98
Row %	4.27	21.87	68.01	5.84	
Column %	2.9	21.71	68.68	24.61	
≥HSIL					
Frequency	14	98	117	204	433
%	0.28	1.98	2.36	4.12	8.75
Row %	3.23	22.63	27.02	47.11	
Column %	0.71	7.29	8.85	64.35	
Total					
Frequency	1964	1345	1322	317	4948
%	39.69	27.18	26.72	6.41	100.00
Colposcopic Biopsy					
Negative					
Frequency	622	16	27	20	685
%	27.81	0.72	1.21	0.89	30.62
Row %	90.80	2.34	3.94	2.92	
Column %	53.57	20.51	5.52	3.93	
ASCUS					
Frequency	142	18	17	7	184
%	6.35	0.80	0.76	0.31	8.23
Row %	77.17	9.78	9.24	3.80	
Column %	12.23	23.08	3.48	1.38	
LSIL					
Frequency	364	33	378	112	887
%	16.27	1.48	16.90	5.01	39.65
Row %	41.04	3.72	42.62	12.63	
Column %	31.35	42.31	77.30	22.00	
≥HSIL					
Frequency	33	11	67	370	481
%	1.48	0.49	3.00	16.54	21.50
Row %	6.86	2.29	13.93	76.92	
Column %	2.84	14.10	13.70	72.69	
Total					
Frequency	1161	78	489	509	2237
%	51.90	3.49	21.86	22.75	100.00

icates moderate agreement, 0.6 to 0.8 shows substantial agreement, and 0.8 to 1.0 is almost perfect agreement.

However, κ values were compared cautiously, for 2 reasons. First, the in-

terpretation of the κ statistic is affected by large differences in disease prevalence.³¹ When disease prevalence is very high or very low (rather than intermediate), the κ values are de-

creased relative to the percentage of agreement, which does not take chance agreement into account. The presentation notes when this might affect interpretation. In general, as measured by the κ statistic, the rates of agreement observed in this ALTS referral population would tend to be lower (relative to percentage of agreement) in a screening population in which disease is rare.³¹ Second, κ statistics vary by the number of diagnostic categories. Hence, κ statistics were computed for 4 × 4 tables and 2 × 2 tables; they cannot be directly compared. For the 4 × 4 tables, the symmetry χ² statistic was used to compare the severity of clinical center vs QC interpretations. Analogously, for the 2 × 2 tables, the McNemar statistic was used.

RESULTS

The primary comparison data for each specimen type are listed in TABLE 1 for monolayer cytology, cervical biopsies, and LEEP specimens, respectively. For each of the 3 data sets, the shaded diagonal represents the proportion of concordant specimens. The boxed data cells indicate the most discordant comparisons. There was only moderate interobserver reproducibility, regardless of specimen type. The κ statistics are compared in TABLE 2. The modest increase in κ values based on weighting suggests that most disagreements were relatively close. There was significant asymmetry in each class of comparison. This suggested a systematic pattern of disagreement between the clinical center and the QC pathologists, with the QC pathologists tending to give less severe interpretations for all 3 types of specimens.

Not surprisingly, the greatest source of disagreement in monolayer cytology results involved ASCUS interpretations (Table 1). Of 1473 original interpretations of ASCUS, the QC reviewer concurred in only 43.0%, rendering less severe readings for most of the rest. Another significant source of variation included HSILs in which concordance was only 47.1%, with 27.0% and 22.6% of the remainder inter-

preted as LSIL or ASCUS by the QC reviewers respectively.

Histologic interpretative reproducibility on biopsies was no better overall than cytologic reproducibility. However, histologic variability derived largely from disagreements about grade CIN1 (including koilocytotic atypia). An interpretation of CIN1 by the clinical center was corroborated by the QC group in only 42.6% of 887 biopsies. Virtually an equal proportion of originally diagnosed CIN1 biopsies (41.0%) were interpreted as negative by the pathology QC group.

An equivocal histologic interpretation (ie, a histologic equivalent of ASCUS) was rarely used, although ALTS is a study in which originally equivocal cytologic interpretations predominate. Most of these problematic cases were due to sample limitations (eg, quality of staining, crush or thermal artifact) that caused difficulty in making subtle distinctions between SIL and normal/reactive. Clinical center pathologists rendered an equivocal histologic interpretation in only 8.2% of 2237 biopsies; similarly, the QC pathologists used an equivocal categorization for biopsy interpretation in only 3.5%. The extremes of interpretation, ie, biopsies categorized as negative or high grade (\geq CIN2), demonstrated good concordance in 90.8% and 76.9% of cases diagnosed by the clinical centers, respectively.

Histologic reproducibility based on LEEP was not better than for other specimen types. Of note, LEEP specimens were much more likely than either cytology or biopsies to represent grades of CIN2 or higher. Despite smaller and skewed numbers of specimens, it was observed that the interpretation of CIN1 was still poorly reproduced, with only 43.8% of original interpretations being corroborated by the QC group.

More condensed 2 × 2 diagnostic tables were composed using each possible binary cutpoint (ie, negative vs \geq ASCUS, \leq ASCUS vs \geq LSIL, and \leq LSIL vs \geq HSIL). The κ statistics for these are shown in TABLE 3. Cytologic interpretations equaling or exceeding

Table 1. Interpretations: Original vs First Quality Control Group Reviewer (cont)*

Original Interpretation	First Quality Control Group Reviewer Interpretation				Total
	Negative	ASCUS	LSIL	\geq HSIL	
LEEP					
Negative					
Frequency	49	2	4	2	57
%	9.16	0.37	0.75	0.37	10.65
Row %	85.96	3.51	7.02	3.51	
Column %	38.89	8.33	4.08	0.70	
ASCUS					
Frequency	17	4	4	2	27
%	3.18	0.75	0.75	0.37	5.05
Row %	62.96	14.81	14.81	7.41	
Column %	13.49	16.67	4.08	0.70	
LSIL					
Frequency	42	10	49	11	112
%	7.85	1.87	9.16	2.06	20.93
Row %	37.50	8.93	43.75	9.82	
Column %	33.33	41.67	50	3.83	
\geqHSIL					
Frequency	18	8	41	272	339
%	3.36	1.50	7.66	50.84	63.36
Row %	5.31	2.36	12.09	80.24	
Column %	14.29	33.33	41.84	94.77	
Total					
Frequency	126	24	98	287	535
%	23.55	4.49	18.32	53.64	100.00

*Shaded diagonals represent concordant interpretations. Boxed data cells indicate the most discordant comparisons. In the shaded diagonals, values in boldface indicate the frequencies of concordance; in the boxed cells, values in boldface indicate the most discordant values. ASCUS indicates atypical squamous cells of undetermined significance; LSIL, low-grade squamous intraepithelial lesion; HSIL, high-grade squamous intraepithelial lesion; and LEEP, loop electrosurgical excision procedure.

Table 2. Summary Data for Original vs First Quality Control Group Diagnosis*

Specimen Type	No.	κ (CI)	Weighted κ † (CI)	Symmetry P Value
Enrollment monolayer	4948	0.46 (0.44-0.48)	0.59 (0.57-0.61)	.001
Colposcopic biopsy	2237	0.46 (0.43-0.49)	0.56 (0.54-0.59)	.001
LEEP	535	0.49 (0.44-0.55)	0.58 (0.52-0.63)	.001

*CI indicates confidence interval; LEEP, loop electrosurgical excision procedure.

†In Table 1, the weights were 1.0 for data cells on the diagonal (ie, exact agreement), 0.67 for cells adjacent to the diagonal, 0.33 for cells 2 units from the diagonal, and 0 for cells 3 units from the diagonal.

Table 3. Summary of Data for Original vs First Quality Control Group Diagnoses, When Divided Into "Disease" vs "Non-Disease" at Different Binary Cutpoints*

Disease Cutpoint	Specimen Type	κ (CI)	McNemar Test P Value
\leq ASCUS vs \geq ASCUS	Enrollment monolayer	0.56 (0.54-0.58)	.001
	Colposcopic biopsy	0.47 (0.44-0.50)	.001
	LEEP	0.46 (0.36-0.55)	.001
\leq ASCUS vs \geq LSIL	Enrollment monolayer	0.64 (0.62-0.67)	.001
	Colposcopic biopsy	0.55 (0.52-0.58)	.001
	LEEP	0.52 (0.44-0.60)	.001
\leq LSIL vs \geq HSIL	Enrollment monolayer	0.51 (0.46-0.55)	.001
	Colposcopic biopsy	0.68 (0.64-0.71)	.08†
	LEEP	0.69 (0.63-0.75)	.001

*CI indicates confidence interval; ASCUS, atypical squamous cells of undetermined significance; LEEP, loop electrosurgical excision procedure; LSIL, low-grade squamous intraepithelial lesion; and HSIL, high-grade squamous intraepithelial lesion.

†The Quality Control Group reviews tended toward less severe interpretations in all comparisons, except for the review of colposcopic biopsies at the cutpoint of \leq LSIL vs \geq HSIL. For this comparison, the original clinical center interpretations tended to be nonsignificantly more severe.

HSIL were uncommon enough to merit caution over comparisons of this κ statistic to others in the table. Cytologic interpretations of LSIL were more reproducible than histologic interpretations of comparable severity. High-grade colposcopy and LEEP biopsy interpretations showed substantial agreement.

Subgroup analyses by individual QC pathologist or by individual clinical center did not significantly alter any of the results or reveal any significant trends over time.

COMMENT

Compared to prior studies, the data available from the ALTS trial are significant for the size of the data set and for the ability to compare common cytologic and histopathologic findings directly. This analysis shows that the interobserver reproducibility of cytologic and histologic interpretations is similar and only moderate. Finer distinctions may be difficult when broader ones are only moderately successful. Nonetheless, an additional study is being pursued to evaluate whether CIN2 can be reproducibly separated from CIN3 within the high-grade group, particularly for histology, where the Bethesda SIL terminology is less accepted. This issue is important for case definition in studies such as the ALTS trial. Obviously, the distinction also has management implications for the minority of clinicians who believe that CIN2 is a reproducible category, not a true cancer precursor, and can be managed differently than CIN3.

For monolayer cytology, the greatest source of variability between the clinical centers and the QC pathologists was the interpretation of ASCUS, which the QC group interpreted as negative in 38.6% of cases. These cytologic smears had a 37% HPV positivity rate much like the HPV positivity rate of 31% among the concordant negative cases (data not shown). Thus, if HPV testing is used as an independent adjudicator of this process, QC-revised interpretations seem likely to be accurate. The other major source of interpretative variability for cytology was the fraction of HSIL clinical center interpretations that were re-

viewed as either LSIL or ASCUS by QC. These represent different problematic sets of cases. Transitions of HSIL to LSIL undoubtedly reflect the difficulty referenced in the literature of trying to separate mild from moderate dysplasia. On the other hand, transitions from HSIL to ASCUS represent the current controversy surrounding small atypical cells of immature metaplastic type and whether these hard-to-interpret cells represent an entity distinct from HSIL.^{10,32}

On biopsy, the κ values were remarkably similar in magnitude to the cytology data. However, the overwhelming source of interpretative variability was the marked tendency of the QC group to review clinical center CIN1 biopsy interpretations as negative. This reflects problems in implementation of criteria for recognizing HPV cytopathic effect in tissue. Although CIN1 is a frequently overcalled interpretation in cervical pathology practice, most of the CIN1 cases reviewed by QC as negative were HPV DNA-positive on the correlated monolayer cytology, suggesting that these disagreements may have been excessive (data not shown). The data for LEEP were similar in this trend. Further direct HPV testing of these samples after microdissection may help clarify the accuracy of the revised interpretations, compared to correlations using HPV tests derived from the temporally related monolayer cytology specimens.

Neither the clinical center nor the QC group was free from important error. The final QC interpretations considered both clinical center and first QC interpretations, as well as additional QC reviews in case of discrepancy. Notably, there was no effect of using the final interpretation on the clinical center agreement rates for LEEP, there was an intermediate improvement on biopsy, and there was the most improvement on cytology (data not shown). However, significant variability was still present and the reasons for these trends are not known.

Two points mitigate the appearance of mediocre reproducibility. First, it is possible that the ALTS population provided slightly heightened diagnostic chal-

lenges, compared with a typical pathology case load. All of the women were referred for a mild cytologic abnormality. Women with easily reproducible, completely negative cytology results or with obviously high-grade results were underrepresented. Secondly, the reproducibility of high-grade colposcopies and LEEP biopsies was substantial, which has important treatment implications. Histologic confirmation of low-grade lesions is more suspect, suggesting that the management of many women is subject to chance. Interestingly, the cytologic diagnosis of LSIL appeared to be more reproducible than the histologic diagnosis. We speculate that these differences are based on the reliability of the criteria applied to individual cells in excellent cytologic preparations compared with the rigor with which these same criteria are applied in histologic sections.

Caveats aside, the data reported in this study probably underestimate the level of variability between groups of pathologists nationally. Most of the pathologists in the trial are academic gynecologic pathologists with a research interest in cervical cancer precursor interpretation and management. In contrast, in many clinical practices, Pap smears are often read in large commercial laboratories whereas biopsies are read locally by community hospital pathologists. Cytohistologic correlation opportunities are decreasing with this unfortunate economic reorganization of cytopathology practice. This problem can potentially be addressed in the future by using the ALTS data set to clarify criteria, revise classification systems, and implement educational tools to help improve interpretative reproducibility within the pathology community. In future works we will focus on whether ASCUS is a useful interpretative entity, what constitutes a CIN1 histologic pattern, and clarification of the variations of HSIL including atypical immature squamous metaplastic cells.

Finally, the need for reproducible interpretations is self-evident. Beyond clinical needs, today's medicolegal environment requires adequate documentation of what is and is not diagnostically pos-

sible. Unrealistic expectations of accuracy, reproducibility, and truth determination fuel many of these malpractice actions. It is possible to achieve moderate to substantial reproducibility in a highly refined environment. However, substantial does not equal perfect and this should provide some basis for understanding and defense in cases based on differences of expert opinion. Indeed, if experts were required to present data on their personal levels of intraobserver and interobserver reproducibility (ideally developed in a standardized objective manner with independent HPV adjudication), then the testimony of many so-called experts would be easier to evaluate.

In this regard, the results of ALTS could stand as a benchmark for the current state of realistic interpretive reproducibility.

Author Contributions: *Study concept and design, acquisition of data, analysis and interpretation of data, drafting of the manuscript, critical revision of the manuscript for important intellectual content, study supervision:* Stoler, Schiffman.

Statistical expertise, obtained funding, administrative, technical, or material support: Schiffman.

Affiliations of the ALTS Group: National Cancer Institute, Bethesda, Md: D. Solomon, M. Schiffman, R. Tarone; **Clinical Centers:** University of Alabama, Birmingham: E. E. Partridge, L. Kilgore, S. Hester; University of Oklahoma, Oklahoma City: J. L. Walker, G. A. Johnson, A. Yadack; Magee-Womens Hospital of the University of Pittsburgh Medical Center Health System, Pittsburgh, Pa: R. S. Guido, K. McIntyre-Seltman, R. P. Edwards, J. Gruss; University of Washington, Seattle: N. B. Kiviat, L. Koutsky, C. Mao, J. M. Haug; **Colposcopy Quality Control Group:** D. Ferris, Medical College of Georgia, Augusta; J. T. Cox,

University of California at Santa Barbara; L. Burke, Beth Israel Deaconess Medical Center Hospital, Boston, Mass; **HPV Quality Control Group:** C. M. Wheeler and C. Peyton-Goodall, University of New Mexico Health Sciences Center, Albuquerque; M. M. Manos, Kaiser Permanente, Oakland, Calif; **Pathology Quality Control Group:** R. J. Kurman, D. L. Rosenthal, and M. E. Sherman, Johns Hopkins Hospital, Baltimore, Md; M. H. Stoler, University of Virginia Health Science Center, Charlottesville; **Cost Utility Analysis Group:** D. M. Harper, Dartmouth Hitchcock Medical Center, Lebanon, NH; **Westat, Coordinating Unit, Rockville, Md:** J. Rosenthal, M. Dunn, J. Quarantillo, D. Robinson; **Information Management Services, Silver Spring, Md:** L. Saxon.

Funding/Support: This work was supported by Public Health Service grants CN55153, CN55154, CN55155, CN55156, CN55157, CN15518, CN55159, and CN55105 from the National Cancer Institute. The following companies have provided support in the form of equipment or supplies at no cost or reduced cost: Cytec Corporation, Boxborough, Mass; DenVu, Tucson, Ariz; Digene, Gaithersburg, Md; National Testing Laboratories, Fenton, Mo; and TriPath Imaging Inc, Elon, NC.

REFERENCES

- Stoler MH. Advances in cervical screening technology. *Mod Pathol.* 2000;13:275-284.
- Confortini M, Biggeri A, Cariaggi MP, et al. Intra-laboratory reproducibility in cervical cytology: results of the application of a 100-slide set. *Acta Cytol.* 1993; 37:49-54.
- Evans DMD, Shelley G, Cleary B, Baldwin Y. Observer variation and quality control of cytodiagnosis. *J Clin Pathol.* 1974;27:945-950.
- Horn PL, Lowell DM, LiVolsi VA, Boyle CA. Reproducibility of the cytologic diagnosis of human papillomavirus infection. *Acta Cytol.* 1985;29:692-694.
- Klinkhamer PJ, Vooijs GP, de Haan AF. Intraobserver and interobserver variability in the diagnosis of epithelial abnormalities in cervical smears. *Acta Cytol.* 1988;32:794-800.
- Lambourne A, Lederer H. Effects of observer variation in population screening for cervical carcinoma. *J Clin Pathol.* 1973;26:564-569.
- Raab SS, Snider TE, Potts SA, et al. Atypical glandular cells of undetermined significance: diagnostic accuracy and interobserver variability using select cytologic criteria. *Am J Clin Pathol.* 1997;107:299-307.
- Raab SS, Geisinger KR, Silverman JF, Thomas PA, Stanley MW. Interobserver variability of a Papanicolaou smear diagnosis of atypical glandular cells of undetermined significance. *Am J Clin Pathol.* 1998;110: 653-659.
- Sherman ME, Schiffman MH, Lorincz AT, et al. Toward objective quality assurance in cervical cytology: correlation of cytopathologic diagnoses with detection of high-risk human papillomavirus types. *Am J Clin Pathol.* 1994;102:182-187.
- Stoler MH. Does every little cell count? don't "ASCUS." *Cancer.* 1999;87:45-47.
- Young NA, Naryshkin S, Atkinson BF, et al. Interobserver variability of cervical smears with squamous-cell abnormalities: a Philadelphia study. *Diagn Cytopathol.* 1994;11:352-357.
- Yobs AR, Plott AE, Hicklin MD, et al. Retrospective evaluation of gynecologic cytodiagnosis, II: interlaboratory reproducibility as shown in rescreening large consecutive samples of reported cases. *Acta Cytol.* 1987;31:900-910.
- Crum CP, Genest DR, Krane JF, et al. Subclassifying atypical squamous cells in Thin-Prep cervical cytology correlates with detection of high-risk human papillomavirus DNA. *Am J Clin Pathol.* 1999;112: 384-390.
- Ferris DG, Cox JT, Burke L, et al. Colposcopy quality control: establishing colposcopy criterion standards for the ALTS trial using cervigrams. *J Lower Genital Tract Dis.* 1998;2:195-203.
- Sellors JW, Nieminen P, Vesterinen E, Paavonen J. Observer variability in the scoring of colpophotographs. *Obstet Gynecol.* 1990;76:1006-1008.
- Cocker J, Fox H, Langley FA. Consistency in the histological diagnosis of epithelial abnormalities of the cervix uteri. *J Clin Pathol.* 1968;21:67-70.
- Creagh T, Bridger JE, Kupek E, Fish DE, Martin-Bates E, Wilkins MJ. Pathologist variation in reporting cervical borderline epithelial abnormalities and cervical intraepithelial neoplasia. *J Clin Pathol.* 1995;48:59-60.
- de Vet HC, Knipschild PG, Schouten HJ, et al. Interobserver variation in histopathological grading of cervical dysplasia. *J Clin Epidemiol.* 1990;43:1395-1398.
- de Vet HC, Knipschild PG, Schouten HJ, et al. Sources of interobserver variation in histopathological grading of cervical dysplasia. *J Clin Epidemiol.* 1992; 45:785-790.
- de Vet HC, Koudstaal J, Kwee WS, Willebrand D, Arends JW. Efforts to improve interobserver agreement in histopathological grading. *J Clin Epidemiol.* 1995;48:869-873.
- Genest DR, Stein L, Cibas E, Sheets E, Zitz JC, Crum CP. A binary (Bethesda) system for classifying cervical cancer precursors: criteria, reproducibility, and viral correlates. *Hum Pathol.* 1993;24:730-736.
- Ismail SM, Colclough AB, Dinnen JS, et al. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *BMJ.* 1989; 298:707-710.
- McCluggage WG, Bharucha H, Caughley LM, et al. Interobserver variation in the reporting of cervical colposcopic biopsy specimens: comparison of grading systems. *J Clin Pathol.* 1996;49:833-835.
- Ringstead J, Amtrup C, Baunsgaard P, et al. Reliability of histo-pathological diagnosis of squamous epithelial changes of the uterine cervix. *Acta Pathol Microbiol Scand.* 1978;86:273-278.
- Robertson AJ, Anderson JM, Beck JS, et al. Observer variation in histopathological reporting of cervical biopsy specimens. *J Clin Pathol.* 1989;42:231-238.
- Siegler EE. Microdiagnosis of carcinoma in situ of the uterine cervix: a comparative study of pathologists' diagnoses. *Cancer.* 1956;9:463-469.
- Seybolt JF, Johnson WD. Cervical cytodiagnostic problems: a survey. *Am J Obstet Gynecol.* 1971;109: 1089-1103.
- Schiffman MH, Adriaana ME, for the ALTS Group. The ASCUS-LSIL Triage Study (ALTS): design, methods, and characteristics of trial participants. *Acta Cytol.* 2000;44:726-742.
- Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd ed. New York, NY: John Wiley & Sons; 1981:218-235.
- Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-174.
- Feinstein AR, Cicchetti DV. High agreement but low kappa, I: the problems of two paradoxes. *J Clin Epidemiol.* 1990;43:543-549.
- Sherman ME, Tabbara SO, Scott DR, et al. "ASCUS, rule out HSIL": cytologic features, histologic correlates, and human papillomavirus detection. *Mod Pathol.* 1999;12:335-342.