

The Interpretive Variability of Cervical Biopsies and Its Relationship to HPV Status

Mark H. Stoler, MD,* Brigitte M. Ronnett, MD,† Nancy E. Joste, MD,‡ William C. Hunt, MA,‡
Jack Cuzick, PhD,§ and Cosette M. Wheeler, PhD,‡|| For the New Mexico HPV Pap Registry
Steering Committee

Abstract: Diagnostic interpretation of a cervical biopsy is a key element in the decision to treat or not to treat a woman with an abnormal screening test. This study assesses the variability of these diagnostic interpretations on a population basis using the New Mexico HPV Pap Registry database. An experienced panel of gynecologic pathologists reviewed a stratified random sample of 6272 biopsies, which was then extrapolated to the entire population of 21,297 biopsies read by the community patholo-

From the *Department of Pathology, University of Virginia, Charlottesville, VA; †Department of Pathology, The Johns Hopkins Medical Institutions, Baltimore, MD; Departments of ‡Pathology; ||Obstetrics and Gynecology, House of Prevention Epidemiology (HOPE), University of New Mexico Health Sciences Center, Albuquerque, NM; and §Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK.

M.H.S., B.M.R., N.E.J. contributed equally.

Conflicts of Interest and Source of Funding: Supported by the National Cancer Institute (NCI) under Award Number R01CA134779 and the National Institute of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health under Award Number U19AI084081 to C.M.W. Roche Molecular Systems Inc. provided HPV Linear Array Genotyping test reagents and equipment to automate HPV genotyping assays through the University of New Mexico. M.H.S. has served as a consultant and/or expert pathologist in HPV-related vaccine and/or diagnostic clinical trials for Merck, Roche, Hologic/Gen-Probe, BD, Qiagen, Cepheid, and Inovio. B.M.R. has served as a consultant and/or expert pathologist in HPV-related vaccine and/or diagnostic clinical trials for Merck. J.C. has received research funding and reagents from Qiagen, Roche, GenProb/Hologic, Abbott, BD, Cepheid, Genera, and Trovagene, and has been personally compensated for Advisory Boards or Speakers Bureau activities from Gen-Probe/Hologic, Abbott, BD, Merck, and Cepheid. C.W. has received equipment and reagents for HPV genotyping from Roche Molecular Systems and contracts from Merck and GSK for HPV vaccine trials through her institution, the University of New Mexico. For the remaining authors none were declared.

Correspondence: Cosette M. Wheeler, PhD, Department of Obstetrics and Gynecology, House of Prevention Epidemiology (HOPE) MSC 02-1670, 1816 Sigma Chi Rd. NE, Albuquerque, NM 87131 (e-mail: cwheeler@salud.unm.edu).

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Website, www.ajsp.com.

Copyright © 2015 Wolters Kluwer Health, Inc. All rights reserved. This is an open-access article distributed under the terms of the Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 (CCBY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

gists. Diagnoses by the community and panel pathologists were compared, and paired diagnoses were correlated with positivity for human papillomavirus 16 (HPV16) and any high-risk HPV as objective measures of progressive potential. Panel agreement with the community diagnosis was 38.2% for cervical intraepithelial neoplasia grade 1 (CIN1), 38.0% for CIN grade 2 (CIN2), 68.0% for CIN grade 3 (CIN3), and 70.6% for cancer. The κ value was 0.46 overall but higher for dichotomous categorization based on CIN2 or CIN3 cutoff points (0.68 and 0.67, respectively). On a population basis, there were fewer CIN1 and more negative diagnoses in the panel review but similar proportions of CIN2 and CIN3. HPV16 and high-risk HPV positivity increased with disease severity, but panel review did not improve the correlation of higher-grade disease with these objective measures. In this population-based study of the variability in cervical diagnoses, we noted significant variability in the community and panel diagnoses, especially for CIN2, the threshold for excisional treatment. New biomarkers are needed to more accurately stratify precursor lesions according to their malignant potential.

Key Words: human papillomavirus prevalence, cervical intraepithelial neoplasia grade 1, 2, 3, cervical cancer, histopathology diagnostic variability, community and expert pathology panel diagnoses, United States

(*Am J Surg Pathol* 2015;39:729–736)

The success of cervical screening relies on several elements, all of which have a subjective or random component. In broad terms, the key elements can be separated by screening sample adequacy, reading of the cytology specimen, colposcopic examination of women with abnormal cytology followed by biopsy of abnormal areas, and interpretation of the biopsy specimen diagnosis. Here we focus on the variability of the diagnostic classification of these biopsies including loop electro-surgical excision procedure specimens.

For most colposcopists, the diagnostic classification of a biopsy is still the major determinant of the decision to treat or not to treat. It has been more than a decade since the National Cancer Institute-led ASCUS and LSIL Triage Study (ALTS) highlighted the issue of diagnostic variability, even among experts, in the interpretation of colposcopically directed biopsies. Criticism of this ALTS

analysis included the fact that all the histopathology interpretations were performed in academic centers, and, thus, critics suggested that the data were not applicable to the real world. But as noted in that paper, the variability in reality may actually be worse. Indeed, other studies including our own^{1–11} have shown consistent results suggesting significant interpathologist variability, especially for certain histologic diagnoses.

Cervical intraepithelial neoplasia grade 2 (CIN2) is the current threshold for treatment, that is, CIN2 or more severe lesions (CIN2+) are treated, except for those diagnosed in younger women, generally under the age of 25 years.^{12,13} Women with less-severe lesions are managed by surveillance.

In ALTS, both CIN1 and CIN2 were found to be highly subjective histologic interpretations.^{14,15} Slightly >50% of CIN1 interpretations made by 1 pathology group were not confirmed by the other. The pattern was the same for CIN2, and both were very different from the more severe and reliable diagnosis of CIN3. Likewise, in the Gardasil vaccine trials, wherein the patients were somewhat younger and lesions even smaller, almost 70% of CIN1 was called negative/normal by expert panel review.

In this study, we analyze, on a population basis, the agreement in CIN classification between the community pathologists and a panel of experienced gynecologic pathologists using the New Mexico HPV Pap Registry (NMHPVPR) database. This interpretative variability is also correlated with 2 objective measures of progressive potential—positivity for human papillomavirus 16 (HPV16) and positivity for any high-risk HPV type.

MATERIALS AND METHODS

Registry

The NMHPVPR acts as a designee of the New Mexico Department of Health and operates under New Mexico Administrative Code (NMAC) 7.4.3.12, which specifies the list of Notifiable Diseases and Conditions for the state of New Mexico. With the intention of monitoring the impact of HPV vaccination, NMAC 7.4.3.12 specified in 2006 that laboratories must report to the NMHPVPR all results of cervical cytology, cervical pathology, and HPV tests performed on New Mexico residents. NMAC 7.4.3.12 was updated in 2009 to include vulvar and vaginal pathology (<http://164.64.110.239/nmac/parts/title07/07.004.0003.htm>). This research study was approved by the University of New Mexico Human Research Review Committee.

Population and Sample Selection

The population consists of women residing in New Mexico who had 1 or more cervical biopsies that were reported to the NMHPVPR during the period 2006 to 2009. Colposcopic procedures with endocervical curettage only were excluded. We restricted the population to women with a biopsy specimen located at 1 of 4 in-state laboratories. These 21,297 women accounted for 77% of all women who had cervical biopsies in New Mexico

during this time period. For women with >1 cervical biopsy during the period, we identified the biopsy with the most severe community diagnosis, and when there was >1 biopsy with this diagnosis, we selected the earliest reported during the sampling period. We attempted to obtain residual blocks from all biopsies with a community diagnosis of CIN2 or greater and aimed to randomly sample 400 CIN1 biopsies and 200 negative biopsies for each of the 4 participating laboratories. The total population numbers and sample of specimens reviewed are given in Table 1 and in greater detail in Supplemental Table 1 (Supplemental Digital Content 1, <http://links.lww.com/PAS/A253>).

The attained sampling fractions for CIN1 and negative biopsies were smaller than targeted because of lack of residual tissue (Supplemental Tables 1 and 2, Supplemental Digital Content 1, <http://links.lww.com/PAS/A253>). Overall, we were able to locate, retrieve, and successfully test for HPV genotypes in 1 or more tissue blocks from 6272 women.

Community Diagnosis

The community diagnosis for each selected tissue block was abstracted from the pathology report as electronically submitted to the NMHPVPR. The abstracted diagnoses were as follows: negative for CIN or malignancy (Negative), CIN grade 1 (CIN1), CIN grade 2 (CIN2), CIN grade 3 (CIN3), carcinoma in situ (CIS), adenocarcinoma in situ (AIS), invasive squamous cell carcinoma (SCC), and invasive adenocarcinoma (ADC). In addition, 5 less-specific categories were occasionally used by community pathologists: CIN1 to CIN2 (CIN1-2), CIN2 to CIN3 (CIN2-3), low-grade lesion without specification of CIN grade (Low-grade [not otherwise specified (NOS)]), high-grade lesion without specification of CIN grade (High-grade [NOS]), and invasive carcinoma without specification of histologic type (Carcinoma [NOS]). These were grouped as stated below by the review panel.

Pathology Panel Adjudicated Diagnosis

The review panel consisted of 3 experienced gynecologic pathologists with academic affiliations (N.E.J., B.M.R., and M.H.S.). One hematoxylin and eosin (H&E) slide from each block was initially reviewed by 2 of the pathologists. When they were in agreement on the histologic diagnosis, no further review was undertaken. When they disagreed on the diagnosis, the slide was sent to the third pathologist for review and final determination of the histologic diagnosis. This occurred for 37% of the cases. In 14% of these cases the third pathologist, who had the final determination, disagreed with both of the first 2 reviews. Each reviewer was blinded to the diagnoses of the other reviewers, the community diagnosis, and all patient information. The diagnostic categories used by the review panel were the same as those used for the community diagnosis as described above, with the exception that CIN3 and CIS were combined, and the categories of

TABLE 1. Sample Counts and Extrapolated Population Counts for Cross-tabulation of Community Versus Panel Diagnosis

	Panel Diagnosis						Total	
	Negative	CIN1	CIN2	CIN3	AIS	SCC		ADC
Community diagnosis								
Negative								
Sample	507	22	11	1	0	0	0	541
Population	7906	390	202	45	0	0	0	8542
CIN1								
Sample	736	539	115	17	3	0	1	1411
Population	3971	3174	682	110	16	0	4	7958
CIN2								
Sample	388	602	872	429	4	0	0	2295
Population	428	668	964	477	5	0	0	2542
CIN3								
Sample	104	68	391	1231	12	5	0	1811
Population	115	77	436	1369	14	5	0	2016
AIS								
Sample	6	0	1	2	38	1	3	51
Population	7	0	1	2	42	1	3	57
SCC								
Sample	2	1	2	33	0	77	1	116
Population	2	1	2	36	0	86	1	129
ADC								
Sample	8	0	0	0	2	1	36	47
Population	9	0	0	0	2	1	40	53
Total								
Sample	1751	1232	1392	1713	59	84	41	6272
Population	12,439	4309	2287	2040	80	94	48	21,297

Numbers in bold are the extrapolated state-wide population counts based on the sampling fractions. Community diagnoses were collapsed into the categories used by the panel as described in the text. Extrapolated population counts may not add up to row and column totals because of rounding error.

CIN1-2, CIN2-3, Low-grade (NOS), High-grade (NOS), and Carcinoma (NOS) were not used.

Tissue Preparation

A “sandwich” technique was used to enable histopathologic review of tissue sections flanking the sections subjected to HPV genotyping as follows: One 4- μ m-thick section was obtained for H&E staining, two 4- μ m-thick sections for HPV genotyping were collected into o-ringed microfuge tubes, a second 4- μ m-thick section was obtained for H&E staining, and, when possible, additional 4- μ m-thick sections were collected onto Fisherbrand Superfrost Plus glass slides for future biomarker evaluations.

HPV Genotyping

Without removal of paraffin wax, the tissue sections cut for HPV genotyping were suspended (50 to 125 μ L) in 10 mM Tris pH 8.0 containing 1 mM EDTA, 0.1% Laureth-12, and 1 mg/mL proteinase K in microfuge tubes and digested with shaking at 65°C for 4 hours, followed by overnight at 37°C. Before polymerase chain reaction-based HPV genotyping, proteinase K was inactivated at 95°C for 15 minutes. Microfuge tubes with the digested sections were centrifuged briefly at 13,000g, whereas the paraffin wax was liquefied and an aqueous-wax interface formed upon cooling. Two and 5 μ L of the aqueous digest from each tissue specimen were used for genotyping with the LINEAR ARRAY HPV Genotyping test (HPV LA; Roche Diagnostics, Indianapolis, IN). The

LINEAR ARRAY HPV Genotyping test is a qualitative test for 37 HPV genotypes incorporating selective polymerase chain reaction amplification with biotinylated PGMY 09/11 L1 region consensus primers and colorimetric detection of amplified products bound to immobilized HPV genotype-specific oligonucleotide probes. Using the Roche LINEAR ARRAY HPV Genotyping test, hybridizations were automated using Tecan Profi-Blot-48 robots (Tecan, Austria) as previously described.¹⁶ The Roche LINEAR ARRAY HPV Genotyping test detects 13 high-risk and 24 low-risk HPV types. HPV52 is not determined directly by a type-specific probe but inferred as previously described.^{16,17} HPV IS39 is a variant of HPV82, and we have classified HPV82 as positive if either probe is positive. Two independent readers interpreted the presence of HPV genotypes using a reference template provided by the manufacturer with >95% agreement. A third reviewer adjudicated discrepancies to render the final HPV genotype results.

Statistical Analysis

This study used a stratified random sample with unequal sampling fractions using 12 sampling strata defined by the community diagnosis (negative, CIN1, and CIN2+) and the 4 laboratories that comprised the study population. To correctly estimate population parameters, it is necessary to account for the varying sample fractions using appropriate weights. Sample fractions and weights are provided in Supplemental Table 1 (Supplemental

Digital Content 1, <http://links.lww.com/PAS/A253>). We used statistical procedures that are appropriate for stratified samples to compute all population proportions, confidence intervals (CIs), and κ values. Variances were computed using Taylor series linearization, and CIs were computed using the logit transformation. SAS (version 9.3) procedure SURVEYFREQ was used to compute proportions. SUDAAN version 11 was used to compute the κ statistics.

For each combination of community diagnosis and panel review diagnosis, we computed the percent positive for HPV16, the percent positive for any high-risk HPV type (HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, or 68), and the percent positive for at least 1 of the 37 HPV types in the LINEAR ARRAY HPV Genotyping test. All were adjusted for the sampling fractions and are, therefore, estimates of the corresponding population proportions. Categories of the community diagnosis that were not used by the panel review were reclassified as follows: Low-grade (NOS) as CIN1, CIN1-2 and High-grade (NOS) as CIN2, CIN2-3 and CIS as CIN3, and Carcinoma (NOS) as SCC.

We computed Cohen κ statistic as a measure of agreement between community and panel review diagnosis. Category-specific κ statistics were computed by dichotomizing the diagnosis at each category (eg, the category-specific κ for CIN2 was computed by classifying the diagnosis as CIN2 and not CIN2). We also computed κ statistics using the dichotomous positive/negative cutoff points of <CIN2 versus CIN2+ and <CIN3 versus CIN3+. We tested for marginal homogeneity in 2x2 tables using the McNemar discordant-pairs odds ratio (OR). A variance was computed for the log OR using the delta method. The CI for the log OR was computed using a normal approximation and then transformed back to the original scale. A Z test on the log OR was used to test for statistical significance. All reported P-values are 2 sided.

We also compared community and panel review diagnosis using a cutoff point of <CIN2 versus CIN2+, stratifying on whether or not the woman had a high-grade cytology in the year previous to the biopsy. High-grade referral cytology was defined as a cytologic interpretation of high-grade squamous intraepithelial lesion (HSIL), atypical squamous cells, cannot rule out HSIL, atypical glandular cells, AIS, or cancer. The κ statistic and the McNemar discordant-pairs OR was computed for each 2x2 table.

For consistency throughout the report, we adopted a convention for paired community and panel review diagnoses of “community diagnosis/panel review diagnosis” (eg, a community diagnosis of CIN3 and a panel review diagnosis of CIN1 will be denoted as CIN3/CIN1).

RESULTS

Table 1 shows the cross-tabulation of the original community diagnosis and the panel review diagnosis for the sample (n = 6272) and when extrapolated to the entire

population (N = 21,297) on the basis of the sampling fractions (Supplemental Table 1, Supplemental Digital Content 1, <http://links.lww.com/PAS/A253>). Negative diagnoses were less common in the community than estimated by panel review (40.1% vs. 58.4%, $P < 0.0001$), whereas CIN1 diagnoses were more common in the community than estimated by panel review (37.4% vs. 20.2%, $P < 0.0001$).

Only 38.2% (539/1411) of CIN1 and 38.0% (872/2295) of community-diagnosed CIN2 were confirmed by panel review (Table 1). On a population basis, the estimated number of CIN2 cases downgraded to CIN1 by the panel (n = 668) was similar to the estimated number of CIN1 cases upgraded by the panel to CIN2 (n = 682). Although the proportion downgraded (26.3%, 668/2542) was greater than the proportion upgraded (8.6%, 682/7958), the absolute numbers were similar because of the much greater prevalence of CIN1 versus CIN2 in the population. Consequently, the percent of the population diagnosed as CIN2 by the community was only slightly greater than that estimated by panel review (11.9% vs. 10.7%, $P = 0.024$).

By comparison, 93.7% (507/541) of negatives, 68.0% (1231/1811) of CIN3, 74.5% (38/51) of AIS, 66.4% (77/116) of SCC, and 76.6% (36/47) of ADC community diagnoses were confirmed by the panel review. A substantial proportion of the SCCs diagnosed by the community were downgraded to CIN3 by the panel (28.4%; 33/116), and 17.0% (8/47) of the ADC diagnoses were called negative by the panel. Very few cases were upgraded to cancer by the panel (5 CIN3 and 1 AIS to SCC; 1 CIN1 and 3 AIS to ADC). Overall, the percent of the population diagnosed as CIN3+ by the community was the same as that estimated by panel review (10.6%).

The corresponding κ values for these comparisons are shown in Table 2. κ values for diagnostic categories dichotomized as CIN2+ versus <CIN2 or CIN3+ versus <CIN3 were higher (0.68 and 0.67, respectively) than the overall κ (0.46). Category-specific κ values were lowest for CIN1 (0.35) and CIN2 (0.32) and higher for negative (0.53), CIN3 (0.64), AIS (0.62), SCC (0.77), and ADC (0.80).

TABLE 2. κ Values for Agreement Between Community Diagnosis and Panel Diagnosis

	κ	95% CI
Overall agreement	0.46	0.43, 0.48
Category-specific agreement		
Negative	0.53	0.50, 0.56
CIN1	0.35	0.31, 0.38
CIN2	0.32	0.29, 0.35
CIN3	0.64	0.61, 0.67
AIS	0.62	0.49, 0.75
SCC	0.77	0.71, 0.84
ADC	0.80	0.69, 0.90
Cutoff points		
CIN2+ vs. <CIN2	0.68	0.65, 0.70
CIN3+ vs. <CIN3	0.67	0.65, 0.70

Among the noncancer diagnoses, increasing severity of the diagnosis by either pathology group correlated with increasing positivity for HPV16 and high-risk HPV (and any HPV, data not shown) (Table 3). ADC showed lower positivity rates, possibly because of contamination with endometrial cancers. In general, diagnoses in which the panel and community pathologists agreed had higher positivity for HPV16 and high-risk HPV than for diagnoses in which one or the other pathology group diagnosed it as less severe, for example, CIN2/CIN2 (41.0%, 90.1%) versus CIN2/CIN1 (22.4%, 84.0%) and CIN1/CIN2 (22.4%, 79.2%). Similar patterns were seen for CIN1 versus normal and CIN2 versus normal, albeit with more variability due to the smaller number of these lesions read by the panel. However, CIN3 lesions downgraded to CIN1 by the panel were much more likely to be HPV positive than CIN1 lesions upgraded to CIN3 (26.6%, 94.3% vs. 0.0%, 48.3%).

There were several diagnoses used by the community that were not used by the pathology panel. These included Low-grade (NOS), CIN1-2, High-grade (NOS), CIN2-3, and CIS. A cross-tabulation of this finer classification against the panel diagnosis is shown in Supplemental Table 2 (Supplemental Digital Content 1, <http://links.lww.com/PAS/A253>) along with an indication of how these were grouped for the main analyses. The majority of Low-grade (NOS) (79.5%, 198/249) and half (50.7%, 35/69) of High-grade (NOS) were called negative on panel review, and this downgrading was higher than when the more specific diagnoses of CIN1, CIN2, and CIN3 were used (46.3%, 15.6%, and 5.6%, respectively). For community-diagnosed CIN1-2, similar proportions were downgraded to CIN1 and upgraded to CIN2 on panel review (35.1% and 32.4%, respectively). For community-diagnosed CIN2-3, more were upgraded to CIN3 than downgraded to CIN2 on panel review (32.6% and 53.3%, respectively).

Supplemental Table 2 (Supplemental Digital Content 1, <http://links.lww.com/PAS/A253>) also shows the results of the HPV genotyping for these finer categories. This broadly supports the grouping of Low-grade (NOS) with CIN1, CIN1-2 with CIN2, High-grade (NOS) with CIN2, and CIN2-3 with CIN3, except that HPV positivity was lower for Low-grade (NOS) versus CIN1 and High-grade (NOS) versus CIN2.

Table 4 shows community versus panel diagnosis using a cutoff point of CIN2 (ie, <CIN2 vs. CIN2+) according to the referral cytology as high-grade versus lower-level abnormalities. Only the community pathologists had access to this information at the time of diagnosis. A significantly greater proportion of cases with high-grade referral cytology was diagnosed as CIN2+ by the community (45.3%) than by the panel (39.8%) (McNemar OR = 2.53; 95% CI = 1.30, 4.91; $P < 0.0001$). When referral cytology was less than high-grade, CIN2+ rates were much lower and similar for the community (16.6%) and the panel (16.5%) (McNemar OR = 1.01; 95% CI = 0.80, 1.28; $P = 0.94$). This clearly indicates an influence of the cytologic diagnosis on the interpretation of histology by community pathologists.

DISCUSSION

Here we describe the level of agreement of routine community diagnosis of cervical specimens to a panel review by expert gynecologic pathologists and also evaluate an objective measurement of progressive potential on the basis of HPV16 and all high-risk HPV positivity. The distribution of disease in the population of 21,297 biopsies from 4 laboratories covering almost 80% of New Mexico's cervical diagnostic pathology was similar to that reported by Galgano et al.¹⁸ The slightly greater fraction of abnormal in the current study may reflect that we selected the worst biopsy diagnosis in a 4-year window of

TABLE 3. Percent HPV16 and High-risk HPV Positivity for Community Versus Panel Diagnosis

Community Diagnosis	HPV Type	Panel Diagnosis							Total
		Negative	CIN1	CIN2	CIN3	AIS	SCC	ADC	
Negative	%HPV16+	4.6	0.0	2.4	—	—	—	—	4.3
	%High-risk HPV+	17.3	31.6	63.1	—	—	—	—	19.5
CIN1	%HPV16+	4.6	9.7	22.4	0.0	—	—	—	8.2
	%High-risk HPV+	25.7	63.4	79.2	48.3	—	—	—	45.7
CIN2	%HPV16+	20.5	22.4	41.0	49.7	—	—	—	34.4
	%High-risk HPV+	62.7	84.0	90.1	93.9	—	—	—	84.6
CIN3	%HPV16+	29.5	26.6	51.3	60.6	65.5	—	—	55.5
	%High-risk HPV+	69.5	94.3	91.1	94.5	100	—	—	92.4
AIS	%HPV16+	—	—	—	—	53.6	—	—	43.6
	%High-risk HPV+	—	—	—	—	100	—	—	93.9
SCC	%HPV16+	—	—	—	76.2	—	55.7	—	60.2
	%High-risk HPV+	—	—	—	100	—	82.6	—	87.4
ADC	%HPV16+	—	—	—	—	—	—	25.4	21.4
	%High-risk HPV+	—	—	—	—	—	—	47.7	44.9
Total	%HPV16+	5.4	11.1	34.1	53.7	53.9	53.3	23.3	14.7
	%High-risk HPV+	22.1	64.3	84.7	92.1	87.8	81.8	48.6	44.6

%HPV16+ is bolded to reflect the main evaluation.
Values are not given for cells with ≤ 5 observations.

TABLE 4. Extrapolated Population Counts for Community versus Panel Diagnosis Stratified by High-grade Referral Cytology

Community Diagnosis	Panel Diagnosis					
	No High-grade Cytology			High-grade Cytology		
	< CIN2	CIN2+	Total	< CIN2	CIN2+	Total
< CIN2	13,176	902	14,078	2265	157	2422
CIN2+	910	1884	2794	397	1606	2003
Total	14,086	2786	16,872	2662	1763	4425
κ (95% CI)	0.61 (0.58, 0.65)			0.74 (0.70, 0.79)		
Discordant-pairs OR (95% CI)	(910/902) = 1.01 (0.80, 1.28)			(397/157) = 2.53 (1.30, 4.91)		

time versus the cross-sectional diagnoses used in Galgano et al.¹⁸

There was a substantial discordance between the grading by the community pathologists and the panel, with an overall κ of 0.46 and substantial disagreement for CIN1 and CIN2. Agreement for CIN3 was better, but the κ was still only 0.67. Some of this disagreement may be due to the fact that the panel read recut tissue sections, which may have differed from the slides read by the community pathologists. In general the slide used by the panel was six to eight 4- μ m-sections below (deeper into the tissue) the community slide. Previous studies have shown a minimal impact of this on diagnosis.^{19,20} To address this further in this study, a subset of 830 cases had the original community slides reread together with the recut tissue sections in a joint review session including all panel members. The projected impact on the results was negligible (data not shown).

Although of only moderate value in assessing progressive potential of precursor lesions,²¹ we have explored the use of HPV positivity, both for the most aggressive HPV type (HPV16) and for all high-risk HPV types, as an objective measure of disease severity. A previous study did not find HPV16 to be a good quality assurance metric,²¹ but this may have been due to the detection of HPV16 in a cervical Pap sample rather than in tissue as reported in this study. Here, clear gradients in HPV positivity with disease severity could be seen across CIN categories, and cases upgraded or downgraded by the panel had higher or lower HPV positivity, respectively. However, when looking at discordant cases, there was no evidence that the panel diagnosis was more correlated with HPV positivity than the community diagnosis. For example, HPV16 and high-risk HPV positivity was similar between CIN1/CIN2 cases (22.4%, 79.2%) and CIN2/CIN1 cases (22.4%, 84.0%), but this was lower than for CIN2/CIN2 cases (41.0%, 90.1%) and higher than for CIN1/CIN1 cases (9.7%, 63.4%).

Other differences between the community diagnosis and the panel were that the community had access to the cytologic diagnosis preceding the tissue biopsy, and this was not available to the panel. This clearly affected the diagnosis, as the community pathologists were more likely than the panel to diagnose CIN2+ when HSIL cytology was present but not when the cytology was of a lower grade. Of note, even in the absence of community-diagnosed CIN2+, HPV16 positivity in specimens with

prior HSIL cytology was greater than that of specimens without HSIL cytology (10.3% vs. 5.5%, $P = 0.068$, data not shown), which has been shown in previous studies.²² As a possibly direct consequence, HPV16 positivity was slightly greater for CIN2+ diagnoses by the community pathologists than by the panel pathology review. The community pathologists may also have used additional immunohistochemical stains (eg, p16 and Ki67) to aid their diagnostic rendering, and this information was not available to the panel pathologists.

It should be recognized that panel review of CIN2 and higher-level community diagnoses was largely complete, whereas smaller sampling fractions were used for CIN1 and negative histology. This led to larger weighting factors for negative community diagnoses, making estimates for some of the discordant cells unreliable. However, for CIN1 a total of 1411 cases were reviewed by the panel, indicating that potential errors due to sampling are small with negligible impact on the conclusions.

Although pathology panel review has been accepted as a research standard for more accurate diagnosis, our data did not support this supposition, and analyses of other cohorts are needed to fully study this issue. Nonetheless, the clinical implications from the above remain clear: many women may be unnecessarily diagnosed as CIN2 and treated, regardless of which diagnosis is considered. Although there is uncertainty and some controversy as to whether excisional treatment leads to increased risk of pregnancy complications such as preterm delivery,^{23,24} it seems plausible that the total volume of tissue removed, especially if treatment is repeated, potentially reduces the competence of the cervix. In addition, unnecessary treatment increases health care costs as does screening these women annually for 20 years.¹³ Yet in the absence of a more definitive way to safely restrict treatment to women who truly need it, the problem of CIN2 will persist as some cases of CIN2 harbor unsampled CIN3 with high malignant potential.²⁵ The key issue is whether precancerous disease is present and therefore should be treated by an ablative or excisional procedure or whether a woman can be safely managed by a conservative approach.

We also noted the use of intermediate and ambiguous diagnostic categories by the community pathologists (eg, CIN1-2, Low-grade [NOS], High-grade [NOS], and CIN2-3). Low-grade (NOS) and High-grade (NOS) categories were substantially downgraded by the panel, and

this was supported by the HPV positivity levels. These diagnoses reflect community practice terminology, and collapsing these categories as we have done does not have a major impact on the discordance between diagnoses. In our opinion, pathologists should assign these diagnoses into accepted standard categories to simplify management decisions and, as needed, to use correlative biomarkers (HSIL or HPV16) or potentially adjunctive stains (eg, p16 immunohistochemistry^{26–28}) to help, rather than continue to imply they can split diagnostic groupings reliably into biologically unsupported categories. Adding additional poorly defined diagnostic subdivisions will only reduce the reliability of histopathology further and, importantly, provides *no clear benefit to patients*. This is especially true for designations that bridge management choices such as CIN1-2, suspicious for CIN, etc. For consistency in communication, our data support using the The Lower Anogenital Squamous Terminology (LAST)²⁷ or the World Health Organization (WHO)-recommended terminology²⁹ as much as possible, which, for WHO, includes using a parenthetical designation of the degree of CIN whenever possible: LSIL (CIN1), HSIL (CIN2), HSIL (CIN3).

In summary, in a large population-based study of community pathologist cervical diagnosis, we observed significant variability in histopathologic diagnosis, for which quality control markers are needed to help guide and refine the diagnosis of CIN, and especially CIN2, which is the clinical threshold for treatment. The use of antecedent cytology results influenced the diagnostic process in a limited way, but more definitive markers are needed. Recognizing the causes of and correlates of diagnostic variability and then applying strategies to decrease it in biologically meaningful ways has direct bearing in a potentially very positive way on patient management. Finally, despite the difficulties with accurate diagnosis of true precancerous cervical lesions, it must be said that cervical screening, diagnosis, and treatment has been highly successful in the prevention of cervical cancer. In our opinion, terminology refinements and the judicious use of tissue-based biomarkers might further help to more accurately and efficiently direct patient management. Toward this end, study slides coupled with unstained slides collected as part of the efforts reported here provide a unique possibility to evaluate the impact of p16 adjudicated histology results in a large US population. Such a study could offer the first population-based estimates of the number of p16 stains that would be used according to the LAST guidelines.²⁷ We plan to estimate the extent to which p16 adjudication reduces the high-grade histology group (HSIL vs. CIN2/3) due to downgrading of p16-negative CIN2 and estimate the extent to which upgrades of CIN1 after p16 use are true missed disease versus overcalled low-grade lesions.

ACKNOWLEDGMENTS

Members of the NMHPVPR Steering Committee reviewed and gave input to the manuscript and supported the concept and directions of the NMHPVPR including the

evaluations presented in this manuscript. The NMHPVPR Steering members participating are as follows: Nancy E. Joste, MD, University of New Mexico Health Sciences Center and Tricore Reference Laboratories, Albuquerque, NM; Walter Kinney, MD, Kaiser Permanente Northern California; Cosette M. Wheeler, PhD, University of New Mexico Health Sciences Center; William C. Hunt, MS, University of New Mexico Health Sciences Center; Alan Waxman, MD, MPH, University of New Mexico Health Sciences Center; David Espey, MD, US Centers for Disease Control and Prevention; Jane McGrath, MD, University of New Mexico Health Sciences Center; Steven Jenison, MD, Community Member; Mark Schiffman, MD, MPH, US National Cancer Institute; Philip E. Castle, PhD, MPH, Global Coalition Against Cervical Cancer, Arlington, VA; Vicki Benard, PhD, US Centers for Disease Control and Prevention; Debbie Saslow, PhD, American Cancer Society; Jane J. Kim, PhD, Harvard School of Public Health; Mark H. Stoler, MD, University of Virginia; Jack Cuzick, PhD, Wolfson Institute of Preventive Medicine, London; Giovanna Rossi Pressley, MSc, Collective Action Strategies, and RWJF Center for Health Policy at University of New Mexico; and Kevin English, RPh, MPH, Albuquerque Area Southwest Tribal Epidemiology Center (AASTEC). No compensation was received for contributions to this manuscript by any named authors or by the NMHPVPR Steering Committee members.

The authors thank the UNM Department of Pathology and HPV Prevention Center team members supporting this very challenging effort conducted over several years including Amanda Pearse, Erika Langsfeld, George Montoya, MaryAnn Jaramillo, Thomas Leete, Lee Fernando, Ann Powell, Teresa Quintana, Carol Morris, Susan Eaton, Michael Robertson, and Scott Horlbeck. We also thank the clinical laboratories that supported this effort including Tricore Reference Laboratories, CHRISTUS St. Vincent Laboratory, SED Laboratories, Quest Diagnostics and Pathology Consultants of New Mexico.

REFERENCES

1. Cai B, Ronnett BM, Stoler M, et al. Longitudinal evaluation of interobserver and intraobserver agreement of cervical intraepithelial neoplasia diagnosis among an experienced panel of gynecologic pathologists. *Am J Surg Pathol*. 2007;31:1854–1860.
2. Creagh T, Bridger JE, Kupek E, et al. Pathologist variation in reporting cervical borderline epithelial abnormalities and cervical intraepithelial neoplasia. *J Clin Pathol*. 1995;48:59–60.
3. de Vet HC, Knipschild PG, Schouten HJ, et al. Interobserver variation in histopathological grading of cervical dysplasia. *J Clin Epidemiol*. 1990;43:1395–1398.
4. 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.
5. Gage JC, Schiffman M, Hunt WC, et al. New Mexico HPV Pap Registry Steering Committee. Cervical histopathology variability among laboratories: a population-based statewide investigation. *Am J Clin Pathol*. 2013;139:330–335.
6. Gage JC, Joste N, Ronnett BM, et al. A comparison of cervical histopathology variability using whole slide digitized images versus glass slides: experience with a statewide registry. *Hum Pathol*. 2013;44:2542–2548.

7. Kato I, Santamaria M, de Ruiz PA, et al. Inter-observer variation in cytological and histological diagnoses of cervical neoplasia and its epidemiologic implication. *J Clin Epidemiol.* 1995;48:1167–1174.
8. Lie AK, Skjeldestad FE, Hagen B, et al. Occurrence of human papillomavirus infection in cervical intraepithelial neoplasia. A retrospective histopathological study of 317 cases treated by laser conization. *APMIS.* 1995;103:693–698.
9. McCluggage WG, Bharucha H, Caughley LM, et al. Inter-observer variation in the reporting of cervical colposcopic biopsy specimens: comparison of grading systems. *J Clin Pathol.* 1996;49:833–835.
10. McCluggage WG, Walsh MY, Thornton CM, et al. Inter- and intra-observer variation in the histopathological reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system. *Br J Obstet Gynaecol.* 1998;105:206–210.
11. Robertson AJ, Anderson JM, Beck JS, et al. Observer variability in histopathological reporting of cervical biopsy specimens. *J Clin Pathol.* 1989;42:231–238.
12. Kinney W, Hunt WC, Dinkelspiel H, et al. Cervical excisional treatment of young women: a population-based study. *Gynecol Oncol.* 2014;132:628–635.
13. Massad LS, Einstein MH, Huh WK, et al. 2012 updated consensus guidelines for the management of abnormal cervical cancer screening tests and cancer precursors. *J Low Genit Tract Dis.* 2013;17:S1–S27.
14. Castle PE, Stoler MH, Solomon D, et al. The relationship of community biopsy-diagnosed cervical intraepithelial neoplasia grade 2 to the quality control pathology-reviewed diagnoses: an ALTS report. *Am J Clin Pathol.* 2007;127:805–815.
15. Stoler MH, Schiffman M. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL Triage Study. *JAMA.* 2001;285:1500–1505.
16. Wheeler CM, Hunt WC, Cuzick J, et al. New Mexico HPV Pap Registry Steering Committee. A population-based study of human papillomavirus genotype prevalence in the United States: baseline measures prior to mass human papillomavirus vaccination. *Int J Cancer.* 2013;132:198–207.
17. Castle PE, Gravitt PE, Solomon D, et al. Comparison of linear array and line blot assay for detection of human papillomavirus and diagnosis of cervical precancer and cancer in the atypical squamous cell of undetermined significance and low-grade squamous intraepithelial lesion triage study. *J Clin Microbiol.* 2008;46:109–117.
18. Galgano MT, Castle PE, Atkins KA, et al. Using biomarkers as objective standards in the diagnosis of cervical biopsies. *Am J Surg Pathol.* 2010;34:1077–1087.
19. Fadare O, Rodriguez R. Squamous dysplasia of the uterine cervix: tissue sampling-related diagnostic considerations in 600 consecutive biopsies. *Int J Gynecol Pathol.* 2007;26:469–474.
20. Stuart LN, Rodriguez AS, Gardner JM, et al. Utility of additional tissue sections in dermatopathology: diagnostic, clinical and financial implications. *J Cutan Pathol.* 2014;41:81–87.
21. Galgano MT, Castle PE, Stoler MH, et al. Can HPV-16 genotyping provide a benchmark for cervical biopsy specimen interpretation? *Am J Clin Pathol.* 2008;130:65–70.
22. Kovacic MB, Castle PE, Herrero R, et al. Relationships of human papillomavirus type, qualitative viral load, and age with cytologic abnormality. *Cancer Res.* 2006;66:10112–10119.
23. Castanon A, Brocklehurst P, Evans H, et al. Risk of preterm birth after treatment for cervical intraepithelial neoplasia among women attending colposcopy in England: retrospective-prospective cohort study. *BMJ.* 2012;345:e5174.
24. Kyrgiou M, Koliopoulos G, Martin-Hirsch P, et al. Obstetric outcomes after conservative treatment for intraepithelial or early invasive cervical lesions: systematic review and meta-analysis. *Lancet.* 2006;367:489–498.
25. Stoler MH, Vichnin MD, Ferenczy A, et al. The accuracy of colposcopic biopsy: analyses from the placebo arm of the Gardasil clinical trials. *Int J Cancer.* 2011;128:1354–1362.
26. Bergeron C, Ordi J, Schmidt D, et al. Conjunctive p16INK4a testing significantly increases accuracy in diagnosing high-grade cervical intraepithelial neoplasia. *Am J Clin Pathol.* 2010;133:395–406.
27. Darragh TM, Colgan TJ, Cox JT, et al. The lower anogenital squamous terminology standardization project for HPV-associated lesions: background and consensus recommendations from the College of American Pathologists and the American Society for Colposcopy and Cervical Pathology. *Arch Pathol Lab Med.* 2012;136:1266–1297.
28. Dijkstra MG, Heideman DA, de Roy SC, et al. p16(INK4a) immunostaining as an alternative to histology review for reliable grading of cervical intraepithelial lesions. *J Clin Pathol.* 2010;63:972–977.
29. Stoler M, Bergeron C, Colgan TJ, et al. Epithelial tumours, part of tumours of the uterine cervix, chapter 7. In: Kurman RJ, Carcangiu ML, Herrington CS, Young RH, eds. *WHO Classification of Tumours of Female Reproductive Organs.* 4th ed. Lyon: IARC; 2014:172–198.