Published in final edited form as:

Am J Surg Pathol. 2010 August ; 34(8): 1077–1087. doi:10.1097/PAS.0b013e3181e8b2c4.

Using Biomarkers as Objective Standards in the Diagnosis of Cervical Biopsies

Mary T. Galgano, MD¹, Philip E. Castle, PhD, MPH², Kristen A. Atkins, MD¹, William K. Brix, MD¹, Sarah R. Nassau, MD¹, and Mark H. Stoler, MD^{1,#}

¹Robert E. Fechner Laboratory of Surgical Pathology, University of Virginia Health System, Charlottesville, VA 22908-0214, USA

²Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, MD 20892, USA

Abstract

Histopathologic diagnosis of cervical biopsies determines clinical management of patients with an abnormal cervical cancer-screening test yet is prone to poor inter-observer reproducibility. Immunohistochemical staining for biomarkers related to the different stages of cervical carcinogenesis may provide objective standards to reduce diagnostic variability of cervical biopsy evaluations but systematic, rigorous evaluations of their potential clinical utility are lacking. To address diagnostic utility of HPV L1, p16^{INK4a}, and Ki-67 immunohistochemical staining for improving diagnostic accuracy, we conducted a community- and population-based evaluation using 1455 consecutive cervical biopsies submitted to the Department of Pathology at the University of Virginia during a period of 14 months. Thin-sections of each biopsy from 1451 of 1455 (99.7%) biopsies underwent evaluation of immunohistochemical stains for three biomarkers, masked to the original diagnosis, and the results were compared to an adjudicated, consensus diagnosis by 3 pathologists. p16^{INK4a} immunostaining, using the strongest staining as the cutpoint, was 86.7% sensitive and 82.8% specific for cervical intraepithelial neoplasia grade 2 or more severe (CIN2+) diagnoses. The p16^{INK4a} performance was more sensitive ($p < 0.001$), less specific ($p < 0.001$), and of similar overall accuracy for CIN2+ compared to the combined performance of all pathologist reviews in routine clinical diagnostic service (sensitivity = 68.9%, specificity = 97.2%). Ki-67 immunostaining was also strongly associated with a CIN2+ diagnosis but its performance at all staining intensities was inferior to p16^{INK4a} immunostaining, and did not increase the accuracy of CIN2+ diagnosis when combined with p16^{INK4a} immunostaining compared to p16^{INK4a} immunostaining alone. We found no utility for L1 immunostaining in distinguishing between CIN and non-CIN. In conclusion, with a rigorous evaluation, we found immunohistochemical staining for p16^{INK4a} to be a useful and reliable diagnostic adjunct for distinguishing biopsies with and without CIN2+.

INTRODUCTION

Cervical cancer prevention programs in the U.S. and other high-resource settings have traditionally relied on the repeated application of a three-stage intervention: 1) screening by

[#]For correspondence: Robert E. Fechner Laboratory of Surgical Pathology, University of Virginia Health System, 1215 Lee St. HEP Rm. 3032, Charlottesville, VA 22908-0214, USA, mhs2e@virginia.edu, ph: 434-982-0284, fax: 434-924-9492.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Pap tests/cervical cytology; 2) colposcopic evaluation (magnified visualization of the cervix after the application of dilute acetic acid) of screen positives and directed biopsy of abnormal-appearing cervical tissue for diagnosis; and 3) excisional or ablative treatment of the cervical tissue in women diagnosed with precancerous lesions. This program of screening, diagnosis, and treatment of precancerous lesions has effectively reduced the burden and mortality due to cervical cancer where it has been effectively implemented.^{1, 15}

Since most screen-positive women do not have clinically important disease, the goal of managing women who screen positive is distinguishing between those women who have precancerous lesions that have malignant potential and need treatment from those who have benign human papillomavirus (HPV)-associated lesions, which are most likely to regress spontaneously and can be monitored without immediate intervention. However, making such distinctions accurately (sensitively and specifically) has proven difficult.

In the U.S., Canada, and some European countries, cervical histopathological diagnoses are currently graded according to the cervical intraepithelial neoplasia (CIN) system: Normal, CIN1, CIN2, CIN3, and cervical cancer. The CIN nomenclature is primarily based on a subjective measure of the thickness, from the stroma towards the surface of the epithelium of the percentage replacement of differentiating epithelial cells by minimally differentiating or proliferating epithelial cells. At least two-thirds of the epithelium is replaced in CIN3, between one-third and two-thirds in CIN2, and one-third or less in CIN1. This classification system was developed before it was known that these intraepithelial lesions, while representing a spectrum of histological changes caused by HPV, are actually two distinct and non-continuous biological processes, one benign and one precancerous.²⁹

In this context, CIN1 is the histopathological manifestation of HPV infection by carcinogenic and non-carcinogenic human papillomavirus (HPV) genotypes. CIN2 and CIN3 (sometimes called carcinoma *in situ*) are considered precancerous diagnoses. However, while CIN3 is considered a definite precancer and is the best surrogate for cancer risk, with 30% of CIN3 in older women becoming invasive over 30 years¹⁹, CIN2 is now considered an equivocal diagnosis of cervical precancer and includes both CIN1/HPV effects as well as some precancerous lesions.⁸ The clinical challenge is to define which is which, i.e. which is of the equivocal diagnoses merits treatment as a precancer.

Histologic assessment of cervical dysplasia is complicated by inter-observer variability equaling that of cytologic interpretation.²⁵ The two key interpretive issues include the distinction of normal from dysplasia (CIN) of any grade; and then benign, mostly transient dysplasia (CIN1) from precancer (CIN3 and some CIN2). The first issue, separating normal from any CIN, is most commonly the result of misclassifying a normal biopsy as CIN1.²⁵ While this error might be viewed as trivial, an erroneous diagnosis of CIN1 provides false assurance to the colposcopist. Specifically if the diagnosis of CIN1 is rendered and indeed the histologic diagnosis is truly normal, the colposcopist thinks a lesion has been found when in fact normal epithelium has been biopsied and thus the ability to make accurate colposcopic and pathologic correlation is severely compromised. Furthermore, while not recommended by current management guidelines²⁸, women diagnosed with CIN1 are sometimes aggressively treated.

Less often, truly precancerous or high-grade lesions are misclassified as negative for dysplasia particularly in the setting of reactive or metaplastic changes or when a biopsy is fragmented or poorly samples the underlying small lesion. Conversely, diagnostic errors resulting in the overcalling of negative as high-grade disease also occur.⁵

Once dysplasia is recognized, the second diagnostic issue is to determine whether the lesion has malignant potential justifying a therapeutic excision, while attempting to minimize the over-treatment of lesions likely to regress. This is an important balance between safety and over-treatment because treatment increases the risk of negative reproductive outcomes for subsequent pregnancies.² Current protocols place this treatment threshold at CIN2²⁸, but CIN2 has been shown to be the least reproducible diagnosis^{4, 8, 25} and has much greater regressive potential than CIN3.^{6, 26} In our view, although CIN2 is a heterogeneous cohort⁸, containing both precancerous and transient lesions, its treatment provides a necessary margin of safety against the risk of cancer.

We have previously proposed a method for reducing inter-observer variability in cervical biopsy interpretations, particularly CIN2, by using fractions attributable to HPV16¹¹; however, widespread genotyping is not yet available, and while applicable as a quality control metric for diagnostic tendencies, it is not helpful in the interpretation of any individual biopsy.

Other biomarkers have been proposed for use in triaging women with cervical dysplasia to increase diagnostic accuracy, akin to incorporating HPV testing into clinical practice as a triage of women with equivocal cytologic abnormalities to determine which women are in need of immediate colposcopy.^{18, 23} Several candidate biomarkers, such as HPV L1, p16^{INK4a}, and Ki-67 as well as others, have been proposed for use on cytologic or histologic specimens based on promising preliminary results.^{9, 13, 16, 17, 20, 30}

To more rigorously evaluate the utility of HPV L1, p16^{INK4a}, and Ki-67 immunohistochemical staining for improving diagnostic accuracy, we conducted a community- and population-based evaluation in almost 1,500 consecutive cervical biopsies. Our goals reflected the diagnostic challenges described above: distinguishing dysplastic from normal tissue and delineating the grade of the lesion among women with dysplasia. This large study has assessed and minimized inter-observer variation by performing a centralized adjudicated review of histology. We also negated any auto-correction of stain interpretations by reviewing hematoxylin and eosin stain (H&E) slides for standard histologic diagnosis and grading the immunohistochemistry results in a masked fashion, thereby permitting an unbiased evaluation of each biomarker's independent contribution to the adjudicated histopathologic assessment.

METHODS

Cervical biopsy review and data collection

Routine clinical H&E-stained slides of all cervical biopsies accessioned for 14 consecutive months (n = 1,455), excluding benign cervical polyps, were prospectively collected with documentation of the routine clinical diagnosis rendered by one of the on-service general anatomic pathologists ("clinical diagnosis"). The median age, mean age, and age range of this population was 26 years, 29 years, and 18-71 years, respectively. The use of the tissues for this study was approved by the institution review board of the University of Virginia, and the analysis of the de-identified data by P.E.C. was determined not to be human subjects research by the NIH Office of Human Subjects Research.

Slides underwent a quality control pathology review blinded to any clinical information except for patient age. A series of biopsies from a single patient were interpreted as separate biopsies. Diagnostic agreement between any two reviewing pathologists provided a "consensus diagnosis". Slides first underwent independent reviews by two pathologists (M.H.S., K.A.A.). If there was diagnostic agreement between the first two reviewers, consensus diagnosis was achieved and no additional reviews were conducted. For cases with

diagnostic disagreement between the first two reviewers, an independent third review was conducted by another pathologist (M.T.G.). If there was diagnostic agreement on two of the three reviews, no additional reviews were performed. Cases with three-way independent disagreement were re-reviewed together, masked to the previous diagnoses, by three reviewing pathologists at a multi-headed microscope until a consensus diagnosis (two-way or three-way agreement) was reached. This consensus was reached by H&E stains only; no immunohistochemical results were used.

All excisional follow-up diagnoses (hysterectomy, cold knife conization, or LEEP) were recorded. More certain diagnostic annotations, such as “favor” or “consistent with”, were considered informative (for example: “squamous dysplasia, favor moderate” was coded as moderate squamous dysplasia). Less certain diagnostic annotations, such as “cannot exclude”, were not considered informative (for example “mild squamous atypia, cannot exclude mild squamous dysplasia” was coded as negative). Biopsies documented as insufficient for diagnosis or diagnosed as benign cervical polyps were excluded from further review.

Results of the preceding Pap tests (the most proximate, available in-house Pap test within 3 years of the colposcopic examination) and Pap tests done at the time of colposcopy (“synchronous Pap test”) were tabulated when available. A subset of 1,359 (93.7%) women had data on either a preceding Pap (n = 790, 54.4%) or a synchronous Pap (n = 980, 67.5%) (411, 28.3%, had both). Any accompanying HPV testing (Hybrid Capture 2; Qiagen, Gaithersburg, MD) was also recorded.

Histology and Immunohistochemistry

All colposcopically- directed cervical biopsies were fixed in 0.25% zinc, neutral-buffered, 10% formalin (Richard-Allan Scientific, Kalamazoo, MI, USA) and embedded in paraffin per routine. These were then cut at 4 μ M for 10 sequential levels and placed on negatively-charged, sialinated slides (Superfrost™, Thermo Fisher Scientific, Inc., Waltham, MA, USA). Slides 1, 2 and 10 were selected for hematoxylin and eosin staining and submitted to the on-service clinical faculty admixed with other routine non-cervical surgical cases. Slides 3-9 were retained for additional analyses, including immunohistochemical staining.

For immunohistochemistry, unstained slides were placed in a 60°C oven for 1 hour, cooled, de-paraffinized, and rehydrated through xylenes and graded ethanol solutions to water. All slides were then treated for 5 minutes in a 3% hydrogen peroxide solution in water to quench endogenous peroxidase. Antigen retrieval for all immunohistochemistry was performed using Target Retrieval Solution (Dako, Carpinteria, CA, USA; product code S1699) in a pressure cooker per the manufacturer’s protocol. Slides were then placed on a Dako Autostainer immunostaining system for use with immunohistochemistry. Non-specific antibody binding was inhibited by incubating sections with serum-free protein block (Dako; product code X0909) for 10 minutes. Primary antibodies against p16^{INK4a} (mtm laboratories AG, Heidelberg, Germany; antibody used per package insert), Ki-67 (Dako; antibody diluted 1:300), and L1 capsid protein of all known HPV types (Cytoactiv Screening Set, Cytoimmun Diagnostics, Pirmasens, Germany; prediluted antibody) were applied for one hour at room temperature. Following washing and incubation with the relevant secondary antibody linked to horseradish peroxidase, also for one hour at room temperature, sections were developed by adding DAB+ (3,3'-Diaminobenzidine) chromogen (Dako; product code K3468) for 5 minutes. Slides were then counterstained in Richard-Allan hematoxylin, dehydrated through graded ethanol solutions, cleared with xylene, and cover-slipped.

Immunohistochemical Scoring

A fourth pathologist (W.K.B.) reviewed and scored all immunohistochemical stains in the epithelium of each biopsy without clinical information or corresponding H&E-stained slides. The scoring of p16 generally included both nuclear and cytoplasmic staining, and was graded as 0 (no staining), 1 (rare singly dispersed cells staining), 2 (patchy but strong staining, often not continuous from basement membrane), and 3 (strong and diffuse staining, usually continuous staining from basement membrane and extending upward in proportion to lesion grade). The scoring of Ki-67 included nuclear staining only, and was scored as 0 (no staining), 1 (1-2 layers of basal/parabasal staining), 2 (diffuse staining confined to the bottom third or superficial staining but with skip areas usually between parabasal and upper zones), and 3 (continuous staining of greater than the lower third of the epithelium). L1 was scored as positive if at least one epithelial cell had discrete nuclear staining. See **Figures 1-5** for examples of typical as well as problematic cases.

Statistical Analyses

A total of 1,451 of 1,455 (99.7%) individual cervical biopsies had both diagnoses and all immunohistochemical staining results. Histologic diagnoses were categorized as follows: Negative, CIN1, CIN2, CIN3/adenocarcinoma *in situ* (AIS), and cancer. We also categorized diagnoses as no CIN (negative) versus CIN (CIN1 or more severe [CIN1+]). To examine the association of the biomarkers with the severity of CIN among women with any CIN, we grouped the small number of cancers with the CIN3/AIS (CIN3/AIS/Cancer).

As measures of diagnostic agreement between the clinical and consensus diagnoses, we calculated the percent total (raw) agreement with binomial 95% confidence intervals (95%CI) and Kappa values with 95%CI, and tested for a tendency for one or the other group to diagnose more severely using the symmetry χ^2 test. Raw agreement, kappa values, and (if appropriate) linearly-weighted kappa values were calculated for the subset of immunohistochemical stained slides read by a second reviewer (K.A.A.) to assess specifically the relative reproducibility of the adjunctive stains.

A Mantel-Haenszel test for trend was used to assess trends of staining positive or intensity with the severity of the consensus diagnosis. Sensitivity, specificity, and Youden's Index (YI) ($YI = \text{sensitivity} + \text{specificity} - 1$), as a metric of accuracy, were calculated for consensus diagnoses of CIN3 and more severe (CIN3+), CIN2+, and CIN1+. McNemar's chi-square was used to test for differences in sensitivity and specificity.

Logistic regression was used to calculate odds ratios (OR) with 95%CI for the association of biomarker immunohistochemical staining with having CIN versus having no CIN, and multinomial logistic regression was used to calculate OR with 95%CI for the association of biomarker immunohistochemical staining with the severity of diagnosis among women with any CIN (CIN3 or CIN2 versus CIN1).

Two-sided Fisher exact tests were used to examine discrepancies in consensus diagnosis and p16^{INK4a} immunohistochemical staining intensity (e.g. <CIN2 and score of 3).

P values of less than 0.05 were considered statistically significant. Stata 8 (College Station, Texas, USA) was used for statistical analyses.

RESULTS

The pair-wise clinical and consensus diagnoses are shown in Table 1. As rendered by academic surgical pathologists in routine general sign-out, there were 755 negative (52.0%), 451 CIN1 (31.1%), 147 CIN2 (10.1%), 92 CIN3/AIS (6.3%), and 6 cancer (0.4%)

diagnoses. The consensus panel review of these biopsies resulted in 748 negative (51.6%), 394 CIN1 (27.2%), 177 CIN2 (12.2%), 127 CIN3/AIS (8.8%), and 5 cancer (0.3%) diagnoses. Raw agreement between the clinical diagnosis and consensus diagnosis was 74.2% (95% = 71.8-76.4%) and the Kappa value was 0.59 (95% CI = 0.55-0.62). There was a 76.8% agreement for negative histology, 49.3% for CIN1, 27.6% for CIN2, 46.0% for CIN3, 83.3% for cancer diagnoses. The consensus review had a tendency to render a more severe diagnosis than the clinical diagnosis ($p < 0.001$); notably the consensus review was more likely to call the clinical diagnosis of CIN2 as CIN3/AIS than the converse (48 versus 20) and to call a community diagnosis of CIN1 as CIN2 than the converse (77 versus 23).

Immunohistochemical staining results for HPV L1, p16^{INK4a}, and Ki-67 were associated/related to each other. Increasing intensity of p16^{INK4a} and Ki-67 immunohistochemical staining were strongly and directly correlated ($p_{\text{trend}} < 0.001$), while HPV L1 positivity decreased with increasing intensity of p16^{INK4a} ($p_{\text{trend}} < 0.001$) and Ki-67 ($p_{\text{trend}} < 0.001$).

Shown in Table 2 are the results of the individual immunohistochemical stains versus the consensus diagnosis. Increasing intensity of p16^{INK4a} ($p_{\text{trend}} < 0.001$) and Ki-67 ($p_{\text{trend}} < 0.001$) immunohistochemical staining was associated with the increasing severity of the consensus diagnosis. Statistically, immunohistochemical staining of HPV L1 ($p_{\text{trend}} < 0.001$) was negatively associated with the increasing severity of the consensus diagnosis. As would be predicted by the natural history of HPV and cervical carcinogenesis, negative histology (96.7%) and cancer (100%) were the most likely to test negative for L1 capsid protein.

We next examined the clinical performance (sensitivity, specificity, and YI) of different positive cutpoints for p16^{INK4a} and Ki-67 staining, and the two markers combined, in relationship to the consensus diagnoses of CIN3+, CIN2+, and CIN1+ (Table 3). Increasing the positive cutpoint for p16^{INK4a} staining increased its specificity and accuracy for CIN3+ and CIN2+ with only a minor decrement in sensitivity due to one case of CIN3/AIS having a staining intensity of 1 (discussed below). By comparison, increasing the positive cutpoint for Ki-67 staining increased its specificity and accuracy for CIN3+ and CIN2+ but there was a greater decrement in sensitivity. Combining the staining results for two biomarkers in general did little to enhance the clinical performance for CIN3+ compared to p16^{INK4a} staining alone: the most accurate combination of p16^{INK4a} (cutpoint = 3) and Ki-67 (cutpoint = 2) was only slightly less sensitive (98.5% vs. 99.2%), more specific (78.1% vs. 74.8%), and had a higher YI (76.6% vs. 74.0%) than p16^{INK4a} (cutpoint = 3) staining alone.

We used logistic regression models to examine the association of each biomarker with having CIN (versus not) and the severity of CIN among those with any CIN (with CIN1 being the reference), mutually adjusting for the other biomarkers. In the models, we used a positive cutpoint of 3 for p16^{INK4a} and Ki-67 staining. HPV L1 (OR = 8.1, 95% CI = 5.1-13.1), p16^{INK4a} (OR = 16, 95% CI = 11-23), and Ki-67 (OR = 18, 95% CI = 4.3-77) staining were associated with having CIN. p16^{INK4a} (OR = 4.7, 95% CI = 3.1-7.3) and Ki-67 (OR = 5.4, 95% CI = 2.9-9.9) staining were associated, and HPV L1 (OR = 0.62, 95% CI = 0.40-0.95) was negatively associated with, having a CIN2+ diagnosis. p16^{INK4a} (OR = 140, 95% CI = 18-1000) and Ki-67 (OR = 22, 95% CI = 11-42) staining were associated, and HPV L1 (OR = 0.18, 95% CI = 0.098-0.35) was negatively associated with having a CIN3+ diagnosis.

Among women with a negative or CIN1 consensus diagnosis, women whose biopsy had p16^{INK4a} immunohistochemical staining score of 3 (see examples in Figure 4) compared to those who had a score of <3 were: 1) more likely to have community diagnosis of CIN1+ (69.6% vs. 25.9%; $p < 0.001$) and CIN2+ (9.7% vs. 1.4%, $p < 0.001$) and 2) more likely to have a preceding or synchronous HSIL Pap (24.6% vs. 16.3%; $p < 0.001$). Among women

with a CIN2 consensus diagnosis, women whose biopsy had p16^{INK4a} immunohistochemical staining score of 3 compared to those who had a score of <3 were: 1) more likely to have community diagnosis of CIN1+ (98.5% vs. 80.0%; $p < 0.001$) and CIN2+ (54.0% vs. 40.0%, $p = 0.2$); 2) more likely to have a preceding or synchronous HSIL Pap (86.2% vs. 68.0%; $p = 0.005$); and 3) more likely to have a diagnosis of CIN2+ on their excised tissue (89.3% vs. 70.3%, $p = 0.03$). The one case of a CIN3 consensus diagnosis with immunohistochemical staining score less than 3 also had a CIN3 clinical diagnosis, a preceding Pap of atypical squamous cells, was positive for HPV L1, and Ki-67 immunohistochemical staining score of 2 (see Figure 5). The subsequent excision was diagnosed as CIN3. This most likely represents a technical failure of the p16 immunostain.

Table 4 compares the sensitivity, specificity, and Youden's Index of p16^{INK4a} with or without Ki-67 immunostaining with individual pathologists' H&E interpretations for the detection of consensus CIN3+ and CIN2+. For an endpoint of CIN3+, p16^{INK4a} with or without Ki-67 immunostaining was an equally sensitive but slightly less specific and less accurate (lower YI) diagnostic method compared to the combined performance of all pathologist reviews (using a diagnostic cutpoint of CIN2 or worse) (Sensitivity, $p = 0.01$; Specificity, $p < 0.001$) and to most of the individual performances of each pathologist singly. Similarly for an endpoint of CIN2+, p16^{INK4a} with or without Ki-67 immunostaining was more sensitive but less specific diagnostic method compared to the combined performance of all pathologist reviews (Sensitivity, $p < 0.001$; Specificity, $p < 0.001$) and to most of the individual performances of each pathologist singly. Ranked according to YI, p16^{INK4a} with or without Ki-67 immunostaining was near the middle of diagnostic accuracy for CIN2+ for the 12 surgical pathologists.

A random subset of immunohistochemical stained slides underwent a second review to assess the reproducibility of grading of the immunochemical stain. The raw agreement and Kappa for HPV L1 immunostaining ($n = 159$) were 96.9% and 0.88, respectively. The raw agreement, Kappa, and linear-weighted Kappa for Ki-67 immunostaining ($n = 162$) were 73.6%, 0.55, and 0.67, respectively. The raw agreement, Kappa, and linear-weighted Kappa for p16^{INK4a} immunostaining ($n = 162$) were 76.5%, 0.64, and 0.80, respectively. Using the p16^{INK4a} immunohistochemical staining score of 3 as the positive cutpoint, the raw agreement was 95.1% and kappa was 0.87. Using the p16^{INK4a} immunostaining score of 3 and a Ki-67 immunostaining of 2+ as the positive cutpoint, the raw agreement was 95.0% and kappa was 0.86. By comparison, the ranges in the raw agreements and kappa values between the three pathologists conducting the consensus reviews were 88.3%-91.9% and 0.67-0.72, respectively.

DISCUSSION

There is little doubt that the pathology community agrees on the utility of adjunctive stains to increase the accuracy of clinical histologic interpretations. Many studies^{9, 12, 13, 16, 17, 20, 30} have provided evidence that Ki-67 and p16^{INK4a} immunohistochemistry are valuable adjunctive aids in the diagnosis of difficult cervical biopsies. However, virtually none of these studies take into account the realities of diagnostic variation and the confounding effect of autocorrelation. Furthermore interpretation of immunohistochemistry can also be subjective and there are no criteria that are widely accepted for what the exact cut point for a positive vs. negative immunohistochemical stain should be. Mix all these variables together and it becomes difficult to be able to get any accurate assessment of a biomarker's performance for its stated purpose.

Patient management is highly contingent on histopathologic pathologic diagnosis. While widely used in clinical studies as the determinative end point for clinical trials, adjudicated

histology is not the standard of care. Furthermore biomarker studies such as the immunochemical stains evaluated in this paper are used adjunctively at the discretion of pathologists. Indeed, part of the decision of whether to apply a special stain or not to a given case is completely dependent on whether the pathologist thinks they have a problem. Given the documented inter-observer variation of histologic diagnosis, together with the fact that these studies have documented that the biggest error in interpretations are at the thresholds of CIN vs. non-CIN as well as even more critically CIN1 vs. CIN2+, one could legitimately ask the questions as to whether biomarkers could be used to override or replace H&E diagnosis. To accomplish this goal, a clear index of performance of each potential marker, alone or in combination, must be established relative to an accepted diagnostic gold standard. This study was designed to address all these issues and the limitations with the existing literature.

We documented again the diagnostic variation in histologic diagnosis as has been previously observed.^{4, 8, 25} While findings in this interdepartmental study, given its relative tendency to under-call CIN2, are slightly different than ALTS^{8, 25}, the overall impression of diagnostic variation and trends are still markedly evident. This was all true despite the fact that many of the faculty have worked together continuously for more than a decade to harmonize diagnoses in order to achieve more consistent management of patients. Indeed, these data from an optimal yet real-world setting highlight the difficulty in achieving reliable diagnostic performance without objective standards, and the circularity of the problem of evaluating biomarkers to address this issue in the absence of a clear gold standard. In other words, these data, like ALTS, likely underestimate severity of the diagnostic challenge. Furthermore, the literature on p16^{INK4a} and Ki-67 is large, but in most cases the sensitivity and specificity estimates provided in that literature are confounded by this lack of a clear gold standard.

In this study, we went to great effort to try to create, if not a gold standard, an excellent reference standard to which we could reliably evaluate these biomarkers. While our reference standard was not perfect, rigorous review and consensus diagnoses does significantly reduce the misclassification of disease endpoints, thereby improving the association of a biomarker to the endpoint of interest.⁷

We demonstrated the following: 1) p16^{INK4a} staining in a strong and diffuse block pattern is highly sensitive for CIN3+ as well as CIN2+ but not CIN1. Thus, p16^{INK4a} immunohistochemical staining is useful in distinguishing high-grade CIN from ≤CIN1 but probably not useful for distinguishing CIN1 from non-CIN (negative); 2) Ki-67 immunohistochemical staining was equally sensitive but less specific for CIN3+ and CIN2+ compared to p16^{INK4a}. The addition of Ki-67 immunostaining to p16^{INK4a} immunostaining, at least on sequential slides did not appreciably improve the diagnostic accuracy for CIN3+ and CIN2+ compared to p16^{INK4a} immunostaining alone; and 3) L1 protein detection, which should be highly correlated with a productive viral infection, was neither sensitive nor specific for any class of cervical neoplasia, most likely because of the complexity of the temporal evolution of the HPV virion production that may be quite transient. Thus our study's original hope that L1 would serve as a reliable marker to distinguish CIN vs. non-CIN was not realized. However, as an aside, finding L1 at the surface of 30% of CIN3 might be evidence of the progression of low-grade lesions to precancer.^{3, 24}

The L1 positive cases that were negative by consensus diagnosis were selectively reviewed (data not shown) and commonly had at least one reviewer diagnosis of CIN1. These clearly represent the difficulty of the negative vs. CIN1 cutpoint that we had hoped to resolve. While L1 may highlight true CIN1 cases missed by our consensus diagnosis, we were disappointed by the overall sensitivity of L1 for consensus CIN1. In our opinion only a

broad spectrum HPV ISH could possibly provide a better gold standard, and as the probes kits improve, we intend to further evaluate these cases with retained sandwiched slides.

Thus, p16^{INK4a} immunohistochemical staining may serve multiple purposes in diagnosing cervical biopsies in routine clinical practice. First, it might be used as a training tool for newly trained pathologists to calibrate their own “receiver-operator characteristic (ROC)” curve to achieve good diagnostic performance for CIN2+. Second, in the same way that HPV DNA testing can be used as an objective QC standard for cytology^{21, 22}, p16^{INK4a} may likewise be used for cervical histopathology. Finally, where there is no or limited pathology review, as in some resource-constrained countries that have implemented a screening, diagnosis, and treatment program for cervical cancer prevention, easily interpreted, reproducible, and strong (3+) p16^{INK4a} immunohistochemical staining might be used in lieu of a pathologist’s review.

While not yet definitively proven and likely to be controversial, given the repeated documented diagnostic variation in H&E interpretation of cervical biopsies, one could envision a reasonable tradeoff in biological and clinical accuracy whereby p16^{INK4a} immunostaining would be a substitute for pathology review at the threshold for high-grade disease requiring treatment. If one considered our consensus diagnosis of CIN 2+ to reflect true disease, treatment based on a p16 score of 3 would appropriately refer 87% (268 of 309) of patients whereas treatment based on the community pathology diagnosis would appropriately refer 69% (213 of 309). Yet, by the same token, using a p16 score of 3 would possibly “over-treat” 39% (155 of 394) of women with consensus CIN1 whereas treatment based on the clinical diagnosis would only over-treat 6% (25 of 394). Thus, using p16^{INK4a} immunostaining would result in a tradeoff of sensitivity for specificity for treatment of precancerous disease compared to a community diagnosis of CIN2+. As noted above, these are conservative estimates of benefits, given the expertise and experience of the pathologists participating in the study.

It is also worth noting that p16^{INK4a}-positive CIN1 may represent a group of women who are at high risk of having or developing CIN2+ for the following reasons. First, we know that colposcopy is inaccurate¹⁴ and some CIN1 may be poorly sampled CIN2+. Second, despite our efforts, there will still be some misclassification of disease by adjudicated, consensus review. While our study demonstrated the value of p16 staining independent from H&E interpretation (specifically avoiding autocorrection), it is reasonable to conclude that for the time being, interpreting the staining result in the context of histology would provide the best diagnosis. Finally, p16^{INK4a}-positive CIN1 may be at significant risk of subsequently developing into CIN2+²⁷ although larger studies to quantify this risk are needed.

Based on our data, treatment based on p16^{INK4a}-positive tissue may be a reasonable tradeoff of sensitivity and specificity for CIN2+, even versus standard H&E interpretation by expert pathologists. Based on our study design it is clear that strong p16 staining is easily interpreted and is highly correlated with the diagnoses based on the adjudicated review. Thus, newly trained or relatively inexperienced practitioners can rely on adjunctive p16 staining for the differential diagnosis of confounding lesions like repair or immature squamous metaplasia. Furthermore, in lower resource settings, where experienced pathologists may not be readily available, the reliable performance of p16^{INK4a} staining may be preferred over H&E interpretation for determining which screen positives need treatment.

In conclusion, in a large community-and population-based series of biopsies that underwent rigorous pathology review, we found immunohistochemical staining for p16^{INK4a} to be a useful and reliable diagnostic adjunct for distinguishing biopsies with and without CIN2+.

Ki-67 immunostaining was inferior to p16INK4a and its inclusion with p16INK4a showed no marked improvement in clinical performance over p16INK4a alone. Thus, we achieved one of two pre-specified goals. Additional research is needed to identify biomarkers useful in distinguishing CIN from non-CIN. In our opinion, the most plausible biomarker, given that all CIN (squamous or glandular) are HPV DNA positive, would be sensitive and specific HPV DNA detection by in situ hybridization.

Acknowledgments

Dr. Castle was supported by the Intramural Research Program of the NIH, National Cancer Institute. Antibodies for the detection of p16INK4a were donated by mtm laboratories AG (Heidelberg, Germany) and for the detection of L1 capsid protein by Cytoimmun Diagnostics (Pirmasens, Germany). Dr. Stoler serves as consultant for mtm laboratories. In addition, Dr. Stoler has been a consultant in clinical trial and HPV DNA test development for Third Wave, Hologic, Qiagen, Roche Molecular Systems and Gen-Probe.

References

1. Volume 10: Cervix Cancer Screening. IARC Press; Lyon, France: 2005.
2. Arbyn M, Kyrgiou M, Simoons C, et al. Perinatal mortality and other severe adverse pregnancy outcomes associated BMJ with treatment of cervical intraepithelial neoplasia: meta-analysis. *BMJ* 2008;337:a1284. doi:10.1136/bmj.a1284. [PubMed: 18801868]
3. Baak, JP.; Bean, SM.; Stoler, MH., et al. Cervical precancer (intraepithelial neoplasia) including colposcopy and functional biomarkers. In: Robboy, SJ.; Anderson, MC.; Mutter, G., et al., editors. *Pathology of the Female Reproductive Tract*. Churchill Livingstone; Philadelphia: 2009.
4. Carreon JD, Sherman ME, Guillen D, et al. CIN2 is a much less reproducible and less valid diagnosis than CIN3: results from a histological review of population-based cervical samples. *Int J Gynecol Pathol* 2007;26:441–6. [PubMed: 17885496]
5. Castle PE, Cox JT, Jeronimo J, et al. An analysis of high-risk human papillomavirus DNA-negative cervical precancers in the ASCUS-LSIL Triage Gynecol Study (ALTS). *Obstet Gynecol* 2008;111:847–56.
6. Castle PE, Schiffman M, Wheeler CM, et al. Evidence for Frequent Regression of Cervical Intraepithelial Neoplasia-Grade 2. *Obstet Gynecol* 2009;113:18–25. [PubMed: 19104355]
7. Castle PE, Schiffman M, Wheeler CM, et al. Impact of improved classification on the association of human papillomavirus with cervical precancer. *Am J Epidemiol* 2010;171:155–63. [PubMed: 20007673]
8. 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;127:805–15. [PubMed: 17439841]
9. Conesa-Zamora P, Domenech-Peris A, Orantes-Casado FJ, et al. Effect of human papillomavirus on cell cycle-related proteins p16, Ki-67, Cyclin D1, p53, and ProEx C in precursor lesions of cervical carcinoma: a tissue microarray study. *Am J Clin Pathol* 2009;132:378–90. [PubMed: 19687314]
10. Demeter LM, Stoler MH, Broker TR, et al. Induction of proliferating cell nuclear antigen in differentiated keratinocytes of human papillomavirus-infected lesions. *Hum Pathol* 1994;25:343–8. [PubMed: 7909306]
11. 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. [PubMed: 18550472]
12. Griesser H, Sander H, Walczak C, et al. HPV vaccine protein L1 predicts disease outcome of high-risk HPV+ early squamous dysplastic lesions. *Am J Clin Pathol* 2009;132:840–5. [PubMed: 19926574]
13. Huang MZ, Li HB, Nie XM, et al. An analysis on the combination expression of HPV L1 capsid protein and p16(INK4a) in cervical lesions. *Diagn Cytopathol*. 2009
14. Jeronimo J, Schiffman M. Colposcopy at a crossroads. *Am J Obstet Gynecol* 2006;195:349–53. [PubMed: 16677597]
15. Kitchener HC, Castle PE, Cox JT. Chapter 7: Achievements and limitations of cervical cytology screening. *Vaccine* 2006;24(Suppl 3):S63–S70. S63-70.

16. Klaes R, Benner A, Friedrich T, et al. p16INK4a immunohistochemistry improves interobserver agreement in the diagnosis of cervical intraepithelial neoplasia. *Am J Surg Pathol* 2002;26:1389–99. [PubMed: 12409715]
17. Liang J, Mittal KR, Wei JJ, et al. Utility of p16INK4a, CEA, Ki67, P53 and ER/PR in the differential diagnosis of benign, premalignant, and malignant glandular lesions of the uterine cervix and their relationship with Silverberg scoring system for endocervical glandular lesions. *Int J Gynecol Pathol* 2007;26:71–5. [PubMed: 17197900]
18. Manos MM, Kinney WK, Hurley LB, et al. Identifying women with cervical neoplasia: using human papillomavirus DNA testing for equivocal Papanicolaou results. *JAMA* 1999;281:1605–10. [PubMed: 10235153]
19. McCredie MR, Sharples KJ, Paul C, et al. Natural history of cervical neoplasia and risk of invasive cancer in women with cervical intraepithelial neoplasia 3: a retrospective cohort study. *Lancet Oncol* 2008;9:425–34. [PubMed: 18407790]
20. Sarian LO, Derchain SF, Yoshida A, et al. Expression of cyclooxygenase-2 (COX-2) and Ki67 as related to disease severity and HPV detection in squamous lesions of the cervix. *Gynecol Oncol* 2006;102:537–41. [PubMed: 16537091]
21. Scott DR, Hagmar B, Maddox P, et al. Use of human papillomavirus DNA testing to compare equivocal cervical cytologic interpretations in the United States, Scandinavia, and the United Kingdom. *Cancer* 2002;96:14–20. [PubMed: 11836698]
22. Sherman ME, Schiffman MH, Lorincz AT, et al. Toward objective quality assurance in cervical cytopathology. Correlation of cytopathologic diagnoses with detection of high-risk human papillomavirus types. *Am J Clin Pathol* 1994;102:182–7. [PubMed: 8042586]
23. Solomon D, Schiffman M, Tarone R. Comparison of three management strategies for patients with atypical squamous cells of undetermined significance: baseline results from a randomized trial. *J Natl Cancer Inst* 2001;93:293–9. [PubMed: 11181776]
24. Stoler MH. Human papillomaviruses and cervical neoplasia: a model for carcinogenesis. *Int J Gynecol Pathol* 2000;19:16–28. [PubMed: 10638450]
25. Stoler MH, Schiffman M. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL Triage Study. *JAMA* 2001;285:1500–5. [PubMed: 11255427]
26. Trimble CL, Piantadosi S, Gravitt P, et al. Spontaneous regression of high-grade cervical dysplasia: effects of human papillomavirus type and HLA phenotype. *Clin Cancer Res* 2005;11:4717–23. [PubMed: 16000566]
27. Wang SS, Trunk M, Schiffman M, et al. Validation of p16INK4a as a marker of oncogenic human papillomavirus infection in cervical biopsies from a population-based cohort in Costa Rica. *Cancer Epidemiol Biomarkers Prev* 2004;13:1355–60. [PubMed: 15298958]
28. Wright TC Jr, Massad LS, Dunton CJ, et al. 2006 consensus guidelines for the management of women with cervical intraepithelial neoplasia or adenocarcinoma in situ. *J Low Genit Tract Dis* 2007;11:223–39. [PubMed: 17917567]
29. Wright TC Jr, Schiffman M. Adding a test for human papillomavirus DNA to cervical-cancer screening. *N Engl J Med* 2003;348:489–90. [PubMed: 12571255]
30. Zhang Q, Kuhn L, Denny LA, et al. Impact of utilizing p16INK4A immunohistochemistry on estimated performance of three cervical cancer screening tests. *Int J Cancer* 2007;120:351–6. [PubMed: 17066437]

Figure 1.

An example of a low-grade SIL/CIN1, that on H&E (A), could be difficult to distinguish from normal or NIL. This lesion is p16 negative (B), possibly due to infection with a low-risk HPV type (genotyping not available). However note the gap in Ki-67 induction between the basal and suprabasal zones (C) often seen in low-grade lesions^{10, 24}, and the clear presence of L1 capsid expression in the superficial zone (D). Magnification on the film plane =15 X in panel B and 30 X in panels A,C, & D.

Figure 2.

A lesion that was problematic due to somewhat tangential sectioning was variably called CIN1 as well as CIN 2 on H&E (A). Typical 3+ p16 positivity in a continuous from the bottom up pattern indicative of high-risk HPV E7 induction of p16 expression (B). An ongoing question is whether this pattern is biologically always indicative of HSIL (see Discussion) Panel C Ki-67 and Panel D L1 expression. Magnification on the film plane = 30X, A-D.

Figure 3.

Typical CIN3 on H&E (A), with strong 3+ p16 and Ki-67 staining (B&C). Note that despite the lack of differentiation there are some “residual” L1 positive nuclei at the surface (D), an unusual but intriguing finding possibly suggesting this lesion evolved from a lower grade lesion (see Discussion). Magnification on the film plane = 30X, A-D.

Figure 4.

Consensus $\leq$ CIN1 on H&E (A), with strong 3+ p16 staining (B). On retrospective review, the focus interpreted as inflamed squamous metaplasia was reconsidered by all reviewers to be missed CIN3 even though the excisional specimen (LEEP) was negative for dysplasia. Consensus $\leq$ CIN1 on H&E (C) with strong 3+ p16 staining (D). On retrospective review, the small free-floating focus has unknown significance, and no excisional specimen was submitted. Magnification on the film plane = 6X in A&B and 15X in C&D

Figure 5. Unanimous consensus CIN3 in lower right of each panel on H&E (A), with negative p16 (B), Ki-67 score=2 (C), and positive L1 (D) albeit mostly in the lower grade (koilocytotic) fragment (upper left of each panel). The excisional specimen (LEEP) also contained CIN3, and all reviewers consider this a likely technical failure of p16 immunostaining. Magnification on the film plane = 6X, A-D

Table 1

A comparison of the community diagnosis and a consensus diagnosis of cervical biopsies. The shadowed cells highlight exact agreement between the two diagnoses. CIN = cervical intraepithelial neoplasia; AIS = Adenocarcinoma *in situ*

			Community Diagnosis					Total
			Neg	CIN1	CIN2	CIN3/AIS	Cancer	
Consensus Biopsy Diagnosis	Neg	n	653	88	6	1	0	748
		%row	86.5%	19.5%	4.1%	1.1%	0.0%	
	CIN1	n	90	279	23	2	0	394
		%row	11.9%	61.9%	15.6%	2.2%	0.0%	
	CIN2	n	10	77	70	20	0	177
		%row	1.3%	17.1%	47.6%	21.7%	0.0%	
	CIN3/AIS	n	2	7	48	69	1	127
		%row	0.3%	1.6%	32.7%	75.0%	16.7%	
	Cancer	n	0	0	0	0	5	5
		%row	0.0%	0.0%	0.0%	0.0%	83.3%	1,451
Total			755	451	147	92	6	

Raw agreement between the clinical diagnosis and consensus diagnosis was 74.2% (95% CI = 71.8-76.4%) and the Kappa value was 0.59 (95% CI = 0.55-0.62).

Table 2

Intensity of p16^{INK4a}, Ki-67, and HPV L1 immunohistochemical staining in relationship to the consensus diagnosis. p16^{INK4a} and Ki-67 staining were scored from 0 (no staining) to 3 (most intense); HPV L1 staining was 0 (negative) or 1 (positive)

			p16 ^{INK4a}				Ki-67				L1		Total
			0	1	2	3	0	1	2	3	Neg	Pos	
Consensus Biopsy Diagnosis	Neg	n	529	125	53	41	94	481	171	2	723	25	748
		%row	70.7%	16.7%	7.1%	5.5%	12.6%	64.3%	22.9%	0.3%	96.7%	3.3%	
	CIN1	n	129	36	74	155	15	129	231	19	268	126	394
		%row	32.7%	9.1%	18.8%	39.3%	3.8%	32.7%	58.6%	4.8%	68.0%	32.0%	
	CIN2	n	19	7	14	137	5	23	102	47	120	57	177
		%row	10.7%	4.0%	7.9%	77.4%	2.8%	13.0%	57.6%	26.6%	67.8%	32.2%	
	CIN3/AIS	n	0	1	0	126	1	0	48	78	106	21	127
		%row	0.0%	0.8%	0.0%	99.2%	0.8%	0.0%	37.8%	61.4%	83.5%	16.5%	
	Cancer	n	0	0	0	5	0	0	1	4	5	0	5
		%row	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	20.0%	80.0%	100.0%	0.0%	
Total			677	169	141	464	115	633	553	150	1,223	229	1,451

Table 3

Clinical performance of p16^{INK4a} and Ki-67 immunostaining in relationship to consensus diagnoses of cervical intraepithelial neoplasia grade 3 or more severe (CIN3+), CIN2+, and CIN1+. Highlighted with a gray background are the three combinations of p16^{INK4a} and Ki-67 that are most accurate (highest Youden's Index) for consensus diagnosis of CIN2+. Se = sensitivity, Sp = specificity, YI = Youden's Index (Sp + Se - 100%)

Outpoint	CIN3+			CIN2+			CIN1+		
	Se	Sp	YI	Se	Sp	YI	Se	Sp	YI
p16 ≥ 1	100.0%	51.4%	51.4%	93.9%	57.7%	51.6%	79.0%	70.8%	49.8%
p16 ≥ 2	99.2%	64.1%	63.3%	91.3%	71.7%	63.0%	72.7%	87.5%	60.2%
p16 = 3	99.2%	74.8%	74.0%	86.7%	82.8%	69.5%	60.2%	94.5%	54.7%
ki ≥ 1	99.2%	8.7%	7.9%	98.1%	9.6%	7.7%	97.0%	12.6%	9.6%
ki ≥ 2	99.2%	56.6%	55.8%	90.6%	62.9%	53.5%	75.4%	76.8%	52.2%
ki = 3	62.1%	94.9%	57.0%	41.7%	98.2%	39.9%	21.0%	99.7%	20.7%
p16 ≥ 1 and ki ≥ 1	99.2%	52.8%	52.0%	92.9%	59.1%	52.0%	77.7%	72.2%	49.9%
p16 ≥ 1 or ki ≥ 1	100%	7.3%	7.3%	99.0%	8.2%	7.2%	98.3%	11.3%	9.6%
p16 ≥ 2 and ki ≥ 1	98.5%	64.7%	63.2%	90.3%	72.3%	62.6%	71.6%	87.6%	59.2%
p16 ≥ 2 or ki ≥ 1	100%	8.1%	8.1%	99.0%	9.1%	8.1%	98.2%	12.5%	10.7%
p16 = 3 and ki ≥ 1	98.5%	74.9%	73.4%	86.1%	82.9%	69.0%	59.8%	94.5%	54.3%
p16 = 3 or ki ≥ 1	100%	8.5%	8.5%	98.7%	9.5%	8.2%	97.4%	12.6%	10.0%
p16 ≥ 1 and ki ≥ 2	99.2%	67.8%	67.0%	88.0%	75.1%	63.1%	67.2%	88.8%	56.0%
p16 ≥ 1 or ki ≥ 2	100%	40.2%	40.2%	96.4%	45.5%	41.9%	87.2%	58.9%	46.1%
p16 ≥ 2 and ki ≥ 2	98.5%	71.4%	69.9%	87.1%	79.1%	66.2%	64.8%	92.9%	57.7%
p16 ≥ 2 or ki ≥ 2	100%	49.4%	49.4%	94.8%	55.6%	50.4%	83.4%	71.4%	54.8%
p16 = 3 and ki ≥ 2	98.5%	78.1%	76.6%	83.5%	85.9%	69.4%	55.7%	96.3%	52.0%
p16 = 3 or ki ≥ 2	100%	53.3%	53.3%	93.9%	59.9%	53.8%	80.0%	75.1%	55.1%
p16 ≥ 1 and ki = 3	62.1%	95.2%	57.3%	40.8%	98.3%	39.1%	20.5%	99.7%	20.2%
p16 ≥ 1 or ki = 3	100%	51.1%	51.1%	94.8%	57.6%	52.4%	79.5%	70.8%	50.3%
p16 ≥ 2 and ki = 3	62.1%	95.5%	57.6%	40.1%	98.5%	38.6%	19.7%	99.7%	19.4%
p16 ≥ 2 or ki = 3	99.2%	63.4%	62.6%	92.9%	71.4%	64.3%	74.0%	87.5%	61.5%
p16 = 3 and ki = 3	62.1%	95.7%	57.8%	39.8%	98.6%	38.4%	19.5%	99.7%	19.2%
p16 = 3 or ki = 3	99.2%	73.9%	73.1%	88.7%	82.4%	71.1%	61.8%	94.5%	56.3%

Table 4

A comparison of the clinical performance of p16^{INK4a} with or without Ki-67 immunostaining and individual pathologists for the detection of consensus CIN 3+ and CIN2+. A positive cutpoint of CIN2 was used to calculate the performance of the pathologists versus the consensus diagnoses. Se = sensitivity, Sp = specificity, YI = Youden's Index (Sp + Se - 100%)

Cutpoint	CIN3+			CIN2+		
	Se	Sp	YI	Se	Sp	YI
p16 = 3	99.2%	74.8%	74.0%	86.7%	82.8%	69.5%
p16 = 3 and ki ≥ 2	98.5%	78.1%	76.6%	83.5%	85.9%	69.4%
p16 = 3 or ki = 3	99.2%	73.9%	73.1%	88.7%	82.4%	71.1%
All Pathologists	93.2%	90.8%	84.0%	68.9%	97.2%	66.1%
A	100%	82.1%	82.1%	83.9%	93.2%	77.1%
B	100%	81.5%	81.5%	100%	88.0%	88.0%
C	100%	83.6%	83.6%	89.5%	97.8%	87.3%
D	83.3%	95.1%	78.4%	62.5%	100%	62.5%
E	100%	87.5%	87.5%	66.7%	92.9%	59.6%
F	71.4%	93.7%	65.1%	56.5%	100%	56.5%
G	84.2%	96.7%	80.9%	55.0%	100%	55.0%
H	100%	93.6%	93.6%	62.9%	100%	62.9%
I	90.9%	92.1%	83.0%	53.6%	94.8%	48.4%
K	100%	76.3%	76.3%	96.6%	90.9%	87.5%
L	100%	87.0%	87.0%	88.0%	96.6%	84.6%
M	95.0%	93.2%	88.2%	56.3%	97.0%	53.3%
A AS PANELIST	97.7%	83.3%	81.0%	94.8%	95.1%	89.9%
L AS PANELIST	96.2%	90.4%	86.6%	76.1%	98.3%	74.4%