Impact of p16\textsuperscript{INK4a} Immunohistochemistry Staining on Interobserver Agreement on the Diagnosis of Cervical Intraepithelial Neoplasia

Alberto Pacchiarotti, MD,\textsuperscript{1} Simona Galeotti, MSc,\textsuperscript{2} Paola Bellardini, MD,\textsuperscript{3} Francesco Chini, MSc,\textsuperscript{4} Guido Collina, MD,\textsuperscript{5} Paolo Dalla Palma, MD,\textsuperscript{6} Bruno Ghiringhello, MD,\textsuperscript{7} Vincenzo Maccallini, MD,\textsuperscript{8} Fabio Musolino, MSc,\textsuperscript{1} Giovanni Negri, MD,\textsuperscript{9} Roberto Pisa, MD,\textsuperscript{10} Ilaria Sabatucci, MD,\textsuperscript{11} and Paolo Giorgi Rossi, PhD\textsuperscript{2}

From \textsuperscript{1}Lega Italiana per la Lotta contro i Tumori, Sezione Provinciale di Latina, Latina, Italy; \textsuperscript{2}Servizio Interaziendale di Epidemiologia, AUSL Reggio Emilia, Reggio Emilia, Italy; \textsuperscript{3}Screening, Latina Local Health Authority ASL di Latina, Latina, Italy; \textsuperscript{4}Laziosanità, Agenzia di Sanità Pubblica Regione Lazio, Rome, Italy; \textsuperscript{5}UO Anatomia Patologica, AUSL di Bologna, Bologna, Italy; \textsuperscript{6}Ospedale Maggiore, AUSL di Bologna, Bologna, Italy; \textsuperscript{7}UO Anatomia Patologica, OIRM S. Anna, Turin, Italy; \textsuperscript{8}UO Anziale di Screening, ASL 1 Abruzzo, Pescina (AQ), Italy; \textsuperscript{9}Dipartimento di Patologia, Central Hospital Bolzano, Bolzano, Italy; \textsuperscript{10}Servizio Anatomia Patologica, Azienda Ospedaliera S. Camillo-Forlanini, Rome, Italy; and \textsuperscript{11}Department of Gynecologic–Obstetrical and Urologic Sciences, University of Rome, Rome, Italy.

Key Words: p16 protein/p16\textsuperscript{INK4a} (cyclin-dependent kinase inhibitor); Cervical cancer; Mass screening; Immunohistochemistry; Cervical intraepithelial neoplasia; Reproducibility

DOI: 10.1309/AJCPCYWVL61SVKFU

ABSTRACT

Objectives: This study aimed to compare the interobserver Cohen $\kappa$ on H&E staining and on H&E plus p16\textsuperscript{INK4a} staining of all cervical biopsy specimens in a population-based screening program.

Methods: All the colposcopy-guided biopsies generated by the routine screening of 23,258 women aged 25 to 64 years were stained with H&E and H&E plus p16. Biopsy specimens were reviewed by six external experts.

Results: The four diagnoses were available in 441 cases. The interobserver $\kappa$ values were 0.52 (95% confidence interval [CI], 0.45-0.58) and 0.48 (95% CI, 0.42-0.56) with H&E and H&E + p16, respectively, when using a five-group classification (normal, CIN 1, CIN 2, CIN 3, and cancer); adopting a two-group classification ($\leq$CIN 1 and $\geq$CIN 2), the values were 0.75 (95% CI, 0.66-0.82) and 0.70 (95% CI, 0.61-0.79), respectively.

Conclusions: The use of p16 on all cervical biopsy specimens in a screening program showed virtually no effect on reproducibility of the histologic diagnosis.
treatment are actually regressive, and recurrence after treatment is a rare event.8

As an indirect measure of diagnostic accuracy, some authors have proposed9,10 using the interobserver Cohen κ11 if the κ values of concordance between two expert observers on the same p16-assessed samples are higher than those assessed with H&E. This would imply that the diagnostic criteria are less subjective and more reproducible and thus that the final diagnosis resulting from these criteria is probably more accurate.

The usefulness of p16 in the differential diagnosis of cervical intraepithelial neoplasia (CIN) when resolving uncertain morphologies is well recognized.6,12-14 However, it is not clear when it should be used, and only recently has there been an effort in systematically studying the cases in which p16 staining is recommended.15 Furthermore, to our knowledge, no population-based studies have been conducted on the use of p16 as a routine auxiliary method for all histologic specimens in a cervical cancer screening program.

The objective of this study was to evaluate whether p16 in histology is a valid biomarker in improving the accuracy of diagnosis. Because of the lack of a gold standard, we aimed to determine whether the diagnostic concordance of different pathologists increases when using p16 and whether the classification of doubtful cases decreases.

Materials and Methods

Population and Study Design

The study was conducted in the province of Latina in central Italy, with a population of 550,000 in 2010; 160,000 women were in the target age group of the screening program (25-64 years). The province-wide organized screening program, which actively invites all the target population for Pap tests every 3 years, has been in place since 1999. The population base of the study consisted of 23,258 women who were alternately placed on two slides, one of which was stained with hematoxylin and coverslipped. Microscopic evaluation was performed according to the CIN classification (CIN 1, CIN 2, and CIN 3).

Data Analysis and Statistical Methods

A gold standard is not available for histology because histology itself is the final and most accurate diagnosis we can obtain for most cervical lesions. Without a gold standard, however, it is not possible to directly measure test validity.

As an indirect measure of diagnostic accuracy, some authors propose using the Cohen κ: if the agreement between two different pathologists on the same sample is higher using p16 than with H&E alone, this means that diagnoses are more reproducible and, probably, more accurate.

A contingency table and the Cohen κ have been calculated to analyze the concordance between observers and to determine whether p16 improves the diagnostic reproducibility. We present the unweighted κ statistics for a five-category classification (normal, including metaplasia; CIN 1; CIN 2; CIN 3, including adenoscarcinoma in situ; and invasive cancer) and a binary classification (≥CIN 2 and ≤CIN 1); 95% confidence intervals of κ statistics were calculated with a bootstrap method (Stata 11.0, StataCorp, College Station, TX).16

The intraobserver concordance between H&E and H&E plus p16 was analyzed to identify which cases were reclassified and to verify whether p16 staining reduces inconclusive diagnoses (inadequate and CIN not otherwise specified [NOS]) and low reproducible diagnoses (CIN 3).
The study was evaluated by the Direzione Sanitaria of the Latina Local Health Unit, which asked for a slightly modified informed consent for colposcopy-guided biopsies explaining the review process and, given the analogy with routine quality control procedures, waived the study from ethical review board submission.

**Results**

Of the 590 biopsies performed, 53 were excluded because it was not possible to recover the slides for external review, 90 were excluded because one of the external reviewers decided not to participate in the study after slide transmission, and six were excluded because it was impossible to link the slide to the case. Thus, the final data set consisted of 441 women with a mean age of 37 years and a range of 25 to 64 years.

These women underwent colposcopy because their Pap test findings were positive. In more than 80% of the cases, a low-grade squamous intraepithelial lesion (L-SIL) or ASC-US was observed. The overall positive predictive value of colposcopy-guided biopsy was 18.7%, with 8.8% for biopsies in women referred for ASC-US and L-SIL and 68.2% for those referred for abnormal squamous cell suggesting high grade and high-grade squamous intraepithelial lesion.

Cohen κ comparison between the local pathologist and the external reviewers for the diagnoses made with only H&E staining is 0.517 and 0.753 for the five- and two-category classifications, respectively, and . The κ values for diagnoses made using H&E plus p16 immunostaining are extremely similar: 0.483 and 0.700 for the five- and two-category classifications, respectively (Tables 2 and 3). The results do not show differences between the reviewers.

**Table 1**

Distribution of the Cytologic and Histologic Diagnoses According to the Original Diagnosis of the Reviewed Cases

<table>
<thead>
<tr>
<th>Cytology</th>
<th>Negative</th>
<th>CIN 1</th>
<th>CIN 2</th>
<th>CIN 3</th>
<th>Invasive Cancer</th>
<th>Inadequate</th>
<th>Total No. (%)</th>
<th>% PPV for CIN 2+</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC-US</td>
<td>82</td>
<td>13</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>106 (23.6)</td>
<td>10.4</td>
</tr>
<tr>
<td>L-SIL</td>
<td>170</td>
<td>66</td>
<td>17</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>257 (57.2)</td>
<td>8.2</td>
</tr>
<tr>
<td>AGC</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5 (1.1)</td>
<td>20.0</td>
</tr>
<tr>
<td>ASC-H</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9 (2.0)</td>
<td>33.3</td>
</tr>
<tr>
<td>H-SIL</td>
<td>17</td>
<td>4</td>
<td>25</td>
<td>18</td>
<td>1</td>
<td>1</td>
<td>66 (14.7)</td>
<td>68.2</td>
</tr>
<tr>
<td>Carcinoma</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2 (0.4)</td>
<td>100.0</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>50.0</td>
</tr>
<tr>
<td>Inadequate</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2 (0.4)</td>
<td>0.0</td>
</tr>
<tr>
<td>Total</td>
<td>282</td>
<td>83</td>
<td>53</td>
<td>27</td>
<td>3</td>
<td>1</td>
<td>449 (100)</td>
<td>18.7</td>
</tr>
</tbody>
</table>

AGC, abnormal glandular cells; ASC-H, abnormal squamous cell suggesting high grade; ASC-US, atypical squamous cells of undetermined significance; CIN, cervical intraepithelial neoplasia; CIN 2+, cervical intraepithelial neoplasia grade 2 or more severe; H-SIL, high-grade squamous intraepithelial lesion; L-SIL, low-grade squamous intraepithelial lesion; PPV, positive predictive value.

**Table 2**

Contingency Table of Interobserver Agreement on Diagnosis Made With H&E and H&E + p16 Staining Based on a Five-Category Classification

<table>
<thead>
<tr>
<th>Diagnosis Made by Local Pathologist</th>
<th>Negative</th>
<th>CIN 1</th>
<th>CIN 2</th>
<th>CIN 3</th>
<th>Invasive Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>H&amp;E staining</td>
<td>220</td>
<td>38</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>260</td>
</tr>
<tr>
<td>Negative</td>
<td>16</td>
<td>48</td>
<td>13</td>
<td>3</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>CIN 1</td>
<td>4</td>
<td>8</td>
<td>17</td>
<td>20</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>CIN 2</td>
<td>2</td>
<td>0</td>
<td>9</td>
<td>13</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>CIN 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Invasive cancer</td>
<td>242</td>
<td>94</td>
<td>40</td>
<td>37</td>
<td>5</td>
<td>418</td>
</tr>
<tr>
<td>H&amp;E + p16 staining</td>
<td>193</td>
<td>30</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>224</td>
</tr>
<tr>
<td>Negative</td>
<td>37</td>
<td>58</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>119</td>
</tr>
<tr>
<td>CIN 1</td>
<td>2</td>
<td>7</td>
<td>13</td>
<td>17</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>CIN 2</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>20</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>CIN 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Invasive cancer</td>
<td>234</td>
<td>98</td>
<td>43</td>
<td>39</td>
<td>4</td>
<td>418</td>
</tr>
</tbody>
</table>

CIN, cervical intraepithelial neoplasia.
The contingency table of the diagnoses made using H&E and H&E plus p16 by the same pathologist illustrates that p16 staining does not reduce inconclusive diagnoses (inadequate and CIN NOS) and low reproducible diagnoses (CIN 2). Table 5. Of the 94 cases classified as CIN 2 with H&E, 61 maintained this classification, four were reclassified as negative (among these there could be atypical immature metaplasia), 13 as CIN 1, and 16 as CIN 3. With p16, two cases of CIN NOS changed to negative, but one negative case and two CIN 1 cases changed to CIN NOS.

Finally, if diagnoses are grouped into two classes (≤CIN 1 and ≥CIN 2), the two different methods lead to the same conclusion regarding the necessity to treat in 95% of the cases: with p16, 19 cases are reclassified as more severe and 18 as less severe. The intraobserver κ of the two readings was 0.71 and 0.86 for the five- and two-category classifications, respectively Table 6 and Table 7, with individual values ranging from 0.53 to 0.95 and from 0.69 to 1, respectively Table 8.

**Discussion**

In this population-based study, we did not find any difference in concordance among pathologists when diagnoses were made using H&E alone or H&E plus p16. Furthermore, we did not observe any reduction in inconclusive diagnoses (inadequate and CIN NOS) and low reproducible diagnoses (CIN 2).
This conclusion is in contrast with those of previous studies.10,19-25 Bergeron and colleagues19 noted a significant 13% increase in sensitivity for CIN 2+ when p16INK4a-immuno­stained slides were added and interpreted together with the H&E slides; they also found that the κ values significantly improved from 0.566 to 0.749. Gurrola-Díaz and colleagues21 observed an increase in κ values among three pathologists when using a binary classification from 0.36, 0.12, and 0.18 for each pair of observers using H&E to 0.59, 0.70, and 0.59 using p16INK4a stains. Dijkstra and colleagues20 found a similar increase in interobserver agreement among three pathologists from a weighted κ ranging from 0.44 to 0.66 in H&E slides to 0.80 with H&E plus p16. Horn and colleagues22 found an increase in κ values for punch biopsies—from 0.49 to 0.64—when H&E-stained slides were read together with p16INK4a-stained slides. Finally, Sari Aslani et al23 found that p16 increased the interobserver agreement as well as consensus.

Our results are in line with the recent College of American Pathologists and American Society for Colposcopy and Cervical Pathology working group recommendations “against the use of p16 IHC as a routine adjunct to histological assessment of biopsy specimens with morphologic interpretations of negative, -IN1, and -IN3” (wherein IN can be interpreted as CIN for our purposes).15

The difference with previous studies could be the result of two factors: (1) κ in this study is high in the H&E analysis as well, and (2) the study population had a low frequency of cases with uncertain interpretation. In fact, the κ for the binary classification was higher than 0.7, which is considered as having very good accordance and standing in the upper part of the distribution of published studies. Dalla Palma and colleagues9 found an overall κ value of 0.65 for a binary classification of CIN 2+, Stoler and Schiffman26 found a value of 0.68, and Carreon and colleagues18 found a 0.70 κ value between two reviewers from the same center and values from 0.42 to 0.46 between reviewers from different centers. Only Malpica and colleagues27 and Crum and colleagues28 reported κ values higher than 0.75.

This study is population based; therefore, it considers all women screened and thus makes it possible to foresee the real impact that using p16 could have. On the other hand, this
approach makes it difficult to observe rare morphologies for which p16 immunostaining could be more useful. In fact, p16 has been demonstrated to be useful in differential diagnosis of some ambiguous morphologies (ie, glands with immature atypical squamous metaplasia,12-14 which in H&E analysis can be erroneously interpreted as CIN 2). In the current study, four cases were classified as CIN 2 using H&E which, after p16 analysis, were found to be negative, including the atypical immature metaplasia.6

The main limitation of our study was that the pathologists did not interpret the slides under the same circumstances: the local pathologist made the interpretation with H&E plus p16 under the psychological pressure of a clinical diagnosis and used H&E only as a quality control and for study purposes. The external reviewers, on the other hand, performed both interpretations for study purposes, without any psychological pressure and in the opposite time order: first H&E and then H&E plus p16. Although the reviewers came from centers throughout Italy, most had participated in several quality control and concordance studies, thereby increasing the baseline agreement in all conditions.

Finally, this population-based study included about 1 year’s workload of a medium-sized screening program with no selection. Unfortunately, we can only report the results from five of the six reviewers who performed the external review of the slides because the data of the sixth reviewer were lost, and it was impossible to re-evaluate the slides. However, because the slides were assigned to the reviewers according to a random sampling, the final estimates were not biased by the incomplete nature of the data.

Although the usefulness of p16 in resolving uncertain morphologies is well known, the use of p16 on all cervical biopsy specimens in a screening program showed virtually no effect on reproducibility of the histologic diagnosis.

The study was supported by the Lega Italiana per la Lotta contro i Tumori, Sezione Provinciale di Latina, and the Latina Local Health Authority (ASL).

Address reprint requests to Dr Giorgi Rossi: Servizio Interaziendale di Epidemiologia, AUSL Reggio Emilia, Via Amendola 2, 42122, Reggio Emilia, Italy; paolo.giorgirosi@ausl.re.it.

Dr Rossi is the principal investigator of a trial funded by the Italian Ministry of Health for which reagents were obtained at a reduced price or for free from Roche, Hologic-GenProbe, and Qiagen.

References

13. Pinto AP, Schlecht NF, Woo TY, et al. Biomarker (ProEx C, p16(INK4A), and MiB-1) distinction of high-grade squamous intraepithelial lesion from its mimics. Mod Pathol. 2008;21:1067-1074.


