Original Article

Interobserver Reproducibility of Cytologic $p16^{INK4a}$/Ki-67 Dual Immunostaining in Human Papillomavirus-Positive Women

Maria Benevolo, MSc[1]; Elena Allia, MSc[2]; Daniela Gustinucci, MSc[3]; Francesca Rollo, MSc[1]; Simonetta Bulletti, MSc[3]; Elena Cesarini, MSc[3]; Basilio Passamonti, MSc[3]; Maria Rosaria Giovagnoli, MD[4]; Elisabetta Carico, MD[4]; Francesca M. Carozzi, MSc[5]; Alessandra Mongia, MSc[5]; Giulia Fantacci, MSc[5]; Massimo Confortini, MSc[5]; Teresa Rubino, MSc[6]; Cristina Fodero, MSc[6]; Sonia Prandi, MD[6]; Natalina Marchi, MD[7]; Angelo Farruggio, MD[7]; Anna Coccia, MD[2]; Luigia Macrì, MD[2]; Bruno Ghiringhello, MD[2]; Guglielmo Ronco, MD[2]; Emma Bragantini, MD[8]; Enzo Polla, MSc[8]; Vincenzo Maccallini, MD[9]; Giovanni Negri, MD[10]; and Paolo Giorgi Rossi, PhD[6,11]; for the New Technologies for Cervical Cancer Screening 2 (NTCC2) Working Group

BACKGROUND: The accumulation of cyclin-dependent kinase inhibitor 2A ($p16^{ink4a}$) protein in a cell is associated with neoplastic progression in precancerous cervical lesions. Dual staining for $p16^{ink4a}$ and Ki-67 has been proposed as a triage test in cervical cancer screening for women who test positive for human papillomavirus DNA. In this study, interobserver reproducibility of the interpretation of this test was assessed. **METHODS:** Forty-two immunostained, liquid-based cytology slides were divided into 2 sets and were interpreted by 17 to 21 readers from 9 different laboratories, yielding a total of 816 reports. Immunostaining results were classified as positive, negative, inconclusive, or inadequate. After evaluation of the first set of slides and before circulation of the second set, the results were discussed in a plenary meeting. The 10 slides with the most discordant results were evaluated again by selected expert cytopathologists. **RESULTS:** The overall κ value was 0.612 (95% confidence interval [CI], 0.523-0.701), it was higher for the positive and negative categories

Corresponding author: Paolo Giorgi Rossi, PhD, Servizio Interaziendale di Epidemiologia, AUSL Reggio Emilia, via Amendola, 2, 42122, Reggio Emilia, Italy; Fax: (011) 390522335460; paolo.giorgirossi@ausl.re.it

[1]Regina Elena National Cancer Institute, Rome, Italy; [2]Central Cervicovaginal Screening Unit and Center for Cancer Epidemiology and Prevention, Turin, Italy; [3]Laboratory Screening Unit, Local Health Authority-Umbria 1, Perugia, Italy; [4]Cytopathology Unit, St Andrea Hospital, "Sapienza" University, Rome, Italy; [5]Human Papillomavirus Laboratory and Molecular Oncology Unit, Regional Cancer Prevention Laboratory, Institute for Cancer Study and Prevention, Florence, Italy; [6]Institute for Research and Health Care (IRCCS), "Arcispedale S. Maria Nuova" Hospital, Reggio Emilia, Italy; [7]Unit 17, Local Health and Social Care Facility, Este Monselice, Italy; [8]Provincial Health Care Service, Trento, Italy; [9]Department of Pathology, Avezzano, Sulmona, and L'Aquila Local Health Trust-Abruzzo, Avezzano, Italy; [10]Department of Pathology, Bolzano Central Hospital, Bolzano, Italy; [11]Interinstitutional Epidemiology Unit, Reggio Emilia Local Health Trust, Reggio Emilia, Italy.

The following are participants in the New Technologies for Cervical Cancer 2 (NTCC2) Working Group: Regione Lazio (Alessandra Barca and Francesco Quadrino); Regina Elena National Cancer Institute, Rome (Maria Benevolo, Amina Vocaturo, Francesca Rollo, Manuela Scalfari, and Giulia Fabbri); Reggio Emilia Local Health Trust (Paolo Giorgi Rossi and Laura Bonvicini); Institute for Cancer Study and Prevention, Florence (Francesca Maria Carozzi, Karin Andersson, Simonetta Bisanzi, Stefania Capassoni, Massimo Confortini, Carmelina Di Pierro, Giulia Fantacci, Anna Iossa, Marzia Matucci, Paola Mantellini, Alessandra Mongia, GiamPaolo Pompeo, Donella Puliti, and Andrea Baldini); Center for Cancer Epidemiology and Prevention, Turin (Guglielmo Ronco, Raffaella Rizzolo, Anna Gillio-Tos, Laura De Marco, Elena Allia, and Bruno Ghiringhello); Provincial Health Care Service, Trento (Mattia Barbareschi, Paolo Dalla Palma, Salvatore Girlando, Enzo Polla, Teresa Pusiol, Sara Condini, and Emma Bragantini); Local Health Authority-Umbria 1, Perugia (Basilio Passamonti, Daniela Gustinucci, Simonetta Bulletti, Elena Cesarini, Nadia Martinelli, Gabriella Vinti, Graziella Principi, Arturo Fabra, Antonella Lucaccioni, Angela Carlani, and Maria Donata Giaimo); Local Health and Social Care Facility 17, Este Monselice (Maria Gabriella Penon, Natalina Marchi, Angelo Farruggio, and Alessandra Bertazzo); Veneto Oncology Institute (Annarosa Del Mistro, Helena Frayle, Martina Rizzi, and Silvia Gori); Veneto Tumor Registry (Manuel Zorzi); and Veneto Regional Coordinated Screening Oncology (Chiara Fedato and Adriana Montaguti).

The manufacturers of CINtec PLUS and ThinPrep kits did not have any influence on the study design, conduct, data analysis, or the decision to publish data.

Received: August 25, 2016; **Revised:** October 14, 2016; **Accepted:** October 17, 2016

Published online Month 00, 2016 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cncy.21800, wileyonlinelibrary.com

($\kappa = 0.692$ and $\kappa = 0.641$, respectively), and it was almost null for the inconclusive category ($\kappa = 0.058$). Considering only readers from laboratories with documented experience, the κ value was higher ($\kappa = 0.747$; 95% CI, 0.643-0.839) compared with nonexperienced centers ($\kappa = 0.498$; 95% CI, 0.388-0.616). The results were similar in both sets of slides ($\kappa = 0.505$ [95% CI, 0.358-0.642] and $\kappa = 0.521$ [95% CI, 0.240-0.698] for the first and second sets, respectively). Reinterpretation of the slides with the most discordant results did not provide any improvement (first evaluation, $\kappa = 0.616$ [95% CI, 0.384-0.866]; second evaluation, $\kappa = 0.403$ [95% CI, 0.182-0.643]). **CONCLUSIONS:** Dual staining for p16 ^{ink4a} and Ki-67 demonstrated good reproducibility, confirming its robustness, which is a necessary prerequisite for its adoption as a triage test in cervical cancer screening programs that use human papillomavirus DNA as a primary test. ***Cancer Cytopathol* 2016;000:000-000.** *© 2016 American Cancer Society.*

KEY WORDS: cervical cancer; cyclin-dependent kinase inhibitor 2A (p16^{ink4a})/Ki-67; dual immunostaining; human papillomavirus; inter-rater agreement.

INTRODUCTION

It has been demonstrated that human papillomavirus (HPV)-based cervical cancer screening is more effective in reducing disease incidence than Papanicolaou (Pap) test-based screening.[1] However, a triage test for HPV-positive women is recommended to reduce colposcopy referral. To date, cytology, with or without information on HPV type, has been the only recommended triage test.[2-4] The most promising biomarkers for a new generation of triage tests are those linked to activation of the viral oncogenes E6 and E7 and the changes caused in the cell by their activation.[5] The overexpression of cyclin-dependent kinase inhibitor 2A (p16^{ink4a}) is the most studied biomarker among those proposed to date, because robust results have been published from longitudinal studies demonstrating both high cross-sectional sensitivity for high-grade lesions[6-8] and the prospective ability to distinguish infections at very low risk of progression from those with a higher probability of progression over the subsequent 3 years.[9] Recently, a novel assay was introduced for the simultaneous assessment of the expression of p16^{ink4a} and the well known proliferation marker Ki-67 in cervical cytology.[10-13] Under physiologic conditions, the coexpression of these proteins does not occur, because they typically induce opposite effects. Indeed, whereas p16^{ink4a} has antiproliferative activity, triggering cell-cycle arrest, Ki-67 is a proliferation-associated protein detectable only in cycling cells. Therefore, concomitant expression of p16^{ink4a} and Ki-67 within the same cell may serve as an indicator of a deregulated cell cycle caused by HPV oncoproteins and is suggestive of a transforming infection.[14-16]

Because the reproducibility of p16^{ink4a} interpretation in cytologic samples is a matter of concern and the possible p16^{ink4a} positivity observed in nondysplastic cells represents an issue in the evaluation of cytologic samples,[17,18] the new combined p16^{ink4a}/Ki-67 immunostaining assay has been proposed to simplify criteria for positivity and possibly to overcome interpretation problems. Indeed, only cells that simultaneously exhibit both p16^{ink4a} and Ki-67 staining are considered positive independent of morphologic criteria, and a single dual-stained cell is sufficient to score a sample as positive according to the manufacturer's indications. However, despite the stringent criteria for positivity, p16^{ink4a}/Ki-67 interpretation is not automated and somehow operator-dependent. To date only a few studies[19,20] have measured the reproducibility of this test in cytologic samples, and no studies have involved more than 2 laboratories.

The current study was conducted within a large trial on the utility and clinical performance of putative biomarkers for triaging HPV-positive women (New Technologies for Cervical Cancer 2 [NTCC2]; clinicaltrials.gov identifier NCT01837693) and was aimed at evaluating interobserver agreement in the interpretation of p16^{ink4a}/Ki-67–immunostained cervicovaginal samples in HPV-positive women.

MATERIALS AND METHODS

Study Design

In total, 42 liquid-based cervicovaginal slides that were immunostained using the CINtec PLUS kit (Roche Diagnostics, Basel, Switzerland) were included in the study. The slides were divided into 2 sets containing 24 slides and 18 slides, and the 2 sets were split into 2 slots to allow faster circulation among centers. The first set of 24 slides was circulated between June 2013 and May 2014, and the second set of 18 slides was circulated between July 2014 and October 2015.

Nine Italian centers (members of the NTCC2 Working Group) involved in cervical cancer screening and/or cervical cancer research took part in the study. Four of the participating centers have been involved in previous research on p16^{ink4a} or have used p16^{ink4a}/Ki-67 dual staining in a clinical setting (Regina Elena National Cancer Institute, Rome [Rome IRE]; Turin; Florence; and Perugia). Because of the pragmatic nature of the NTCC2 trial, it was essential to assess the reproducibility of the test for both experienced and nonexperienced laboratories. After the first interpretation and a preliminary analysis of the results, a set of the 10 slides with the most discordant reports was formed and reinterpreted by the 4 centers that had documented experience in the use of the test and by an external center (G.N., Central Hospital, Bolzano) that had documented expertise in the field.

Immunostaining

Slides were prepared from liquid-based cytological samples using the ThinPrep 2000 processor (Hologic, Bedford, Mass) and were immunostained using the CINtec PLUS kit (Roche Diagnostics) according to the manufacturer's instructions at 4 of the 9 participating centers (Rome IRE, Perugia, Rome S. Andrea Hospital/"Sapienza" University [Rome SAH], and Turin). All immunostained slides were analyzed and scored independently by 1 to 6 cytopathologists per laboratory who were trained in the interpretation of CINtec PLUS-stained slides and blinded to the cytologic diagnoses.

Interpretation Criteria and Training

Before interpretation of the first set began (May 2013), approximately 15 microscope-projected images were evaluated in a plenary session to establish the criteria for immunostaining interpretation, classification, and reporting of results. This initial session was attended by all readers with a few exceptions. Results from the overall assessment were reported in 4 categories: positive, negative, inadequate, and inconclusive. Samples were scored as positive when double immunoreaction (cytoplasmic and/or nuclear brown staining for p16^{ink4a} together with nuclear red staining for Ki-67) was observed within at least 1 cell. Slides were scored as negative when immunoreactivity was evident for neither marker or for only 1 of the 2 markers.

According to the 2001 Bethesda System,[21] slides that had <5000 well preserved and well visualized squamous epithelial cells were considered inadequate for dual-staining evaluation. However, specimens that had insufficient nucleate squamous cells but exhibited p16^{ink4a}/Ki-67–immunostained cells were reported as satisfactory for evaluation and were recorded as positive.

The slides for which no conclusive interpretation could be reached (for example, because of the presence of nonspecific or weak staining) were scored as inconclusive. Readers also detailed the presence of both p16^{ink4a} and Ki-67 as a single staining and evaluated the cellularity of the slide. Moreover, the presence of ≥ 5 double-stained cells was used as an arbitrary cutoff and was recorded.

After the first set of slides had been reviewed by all centers, a preliminary analysis of the data was performed and the results were presented and discussed in a second plenary meeting in May 2014. The results of the evaluation of the second set were presented and discussed in December 2015. All the readers, with few exceptions, participated in these meetings.

Data Analysis

The proportion of each assessment made is reported for each slide. Kappa (κ) values for multiple readers[22] are reported for overall agreement as well as for specific categories. For an analysis of sensitivity, κ values also are reported excluding the 5 centers that did not use p16^{ink4a}/Ki-67 in their routine practice. Moreover, we analyzed data separately from the 2 sets of slides. For the 10 most discordant slides that were reinterpreted by 6 experienced readers, κ values for 2 readers are reported.[23] Ninety-five percent confidence intervals (CIs) of the overall κ values were calculated using the bootstrap method with bias correction[24] and 1000 simulations. All the analyses were performed using the STATA 13.0 statistical package (Stata Corporation, College Station, Tex).[25] This study was approved by all local ethics committees.

RESULTS

Twenty-five readers in 9 centers were involved in the study, including 1 from Avezzano, 2 from Florence, 3 from Perugia, 6 from Reggio Emilia, 2 from Rome IRE, 2 from Rome SAH, 2 from Trento, 1 from Este Monselice, 5 from Turin, and 1 from Bolzano. However, all 42 slides were interpreted by more than 1 reader at only 4 centers (Florence, Perugia, Reggio Emilia, and Turin). The 2 slots from the first set were evaluated by 20 and 21 readers, respectively, and the 2 slots from the second set were

TABLE 1. Frequency of Results by Slide, Including Re-Evaluation of the 10 Selected Slides With the Most Discordant Results

Slide ID	Negative	Positive	Inconclusive	Inadequate	>5 Positive Cells	Total
1.1	0	19	1	0	18	20
1.2	16	0	4	0	0	20
1.3	0	20	0	0	19	20
1.4	0	20	0	0	0	20
1.5[a]	17	7	3	0	0	27
1.6	0	20	0	0	17	20
1.7[a]	7	15	5	0	13	27
1.8[a]	19	7	1	0	0	27
1.9	0	20	0	0	18	20
1.10	13	0	0	7	13	20
1.11	0	20	0	0	0	20
1.12	2	0	0	18	15	20
2.1	0	21	0	0	21	21
2.2	1	18	2	0	0	21
2.3	0	21	0	0	19	21
2.4	0	21	0	0	18	21
2.5	0	20	1	0	1	21
2.6[a]	7	17	4	0	7	28
2.7	2	18	1	0	9	21
2.8	2	18	1	0	15	21
2.9	0	17	0	4	0	21
2.10[a]	1	13	0	14	2	28
2.11	0	19	0	2	16	21
2.12	1	20	0	0	10	21
3.1	11	1	0	5	0	17
3.2[a]	3	21	0	0	11	24
3.3	17	0	0	0	0	17
3.4	15	2	0	0	2	17
3.5	0	17	0	0	11	17
3.6	15	2	0	0	0	17
3.7	15	1	1	0	0	17
3.8[a]	3	21	0	0	18	24
3.9	15	2	0	0	0	17
4.1[a]	21	2	3	0	0	26
4.2	18	1	0	0	1	19
4.3[a]	22	3	1	0	0	26
4.4	18	0	0	1	0	19
4.5[a]	19	6	0	1	0	26
4.6	17	0	1	1	0	19
4.7	17	1	1	0	1	19
4.8	18	0	0	1	0	19
4.9	0	19	0	0	18	19
Total						886

[a] These slides were selected for the second reading.

evaluated by 17 and 19 readers, respectively. Finally, in total, 816 independent readings were available.

Agreement on the interpretation for each individual slide varied from 100% in 11 slides to a single slide for which 7 readers gave a positive report, 10 gave a negative report, and 4 gave an inconclusive evaluation (Table 1). Six slides were reported as inconclusive by $\geq$2 readers. Six slides were considered inadequate because of insufficient cellularity by $\geq$2 readers.

Slides with low agreement mostly had <5 double-stained cells, and/or presented overlapping cells (Figs. 1C,D), or had weak and/or faded staining (Fig. 1B).

The overall κ value among all readers from the 9 participating laboratories was 0.612 (95% CI, 0.523-0.701). Agreement was higher for the negative and positive categories ($\kappa = 0.641$ and $\kappa = 0.692$, respectively) than for the inadequate category ($\kappa = 0.477$); whereas, for the inconclusive category, the κ value was almost null ($\kappa = 0.058$) (Table 2). When a positivity threshold for slides with >5 double-stained cells was used, the overall κ value was similar ($\kappa = 0.618$; 95% CI, 0.444-0.759).

Overall agreement increased when we took into account only the reports from the 4 centers that had expertise in the use of the test ($\kappa = 0.747$; 95% CI, 0.643-0.839), whereas the κ value was 0.498 (95% CI, 0.388-

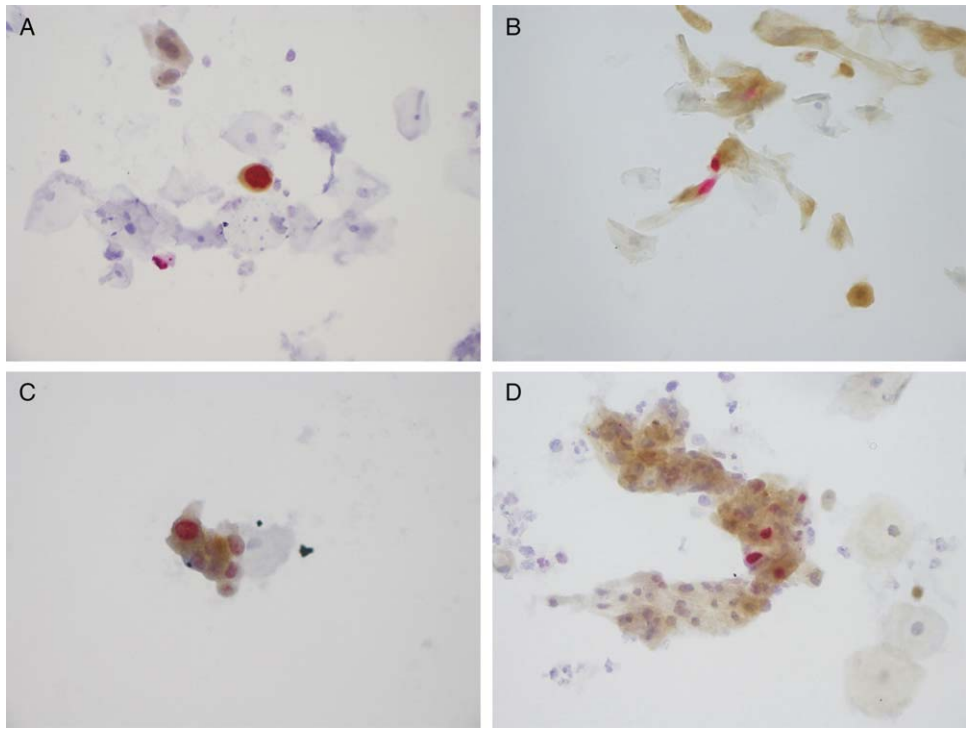

Figure 1. Immunocytochemical double staining of cervicovaginal samples for cyclin-dependent kinase inhibitor 2A (p16^{ink4a})Ki-67 expression is shown. (A) Clearly double-stained cells (unanimously interpreted as positive) are observed. (B,C) Partial p16^{ink4a} staining of the cytoplasm is observed in the presence of Ki-67–positive nuclei from (B) single cells and (C) cell clusters in which cellular borders are not well defined (these were interpreted as positive by the majority of the readers but inconclusive by some others). (D) Possibly aspecific and diffuse p16^{ink4a} staining and some Ki-67–stained nuclei are observed in a cluster of cells (these were interpreted as negative by the majority of readers but inconclusive by some others; counterstaining with hematoxylin, original magnification × 40 in A-D).

TABLE 2. Interobserver κ Values for the Overall Set and Sensitivity Analyses

Variable	No. of Readers (No. of Readings)	κ Value				Overall	
		Negative	Positive	Inconclusive	Inadequate	κ Value	95% CI
All centers	23 (816)	0.641	0.692	0.058	0.477	0.612	0.523-0.701
Only experienced centers	12 (453)	0.786	0.823	0.098	0.523	0.747	0.643-0.839
Only nonexperienced centers	11 (363)	0.520	0.578	0.106	0.426	0.498	0.318-0.616
First set	21 (492)	0.469	0.604	0.070	0.549	0.505	0.358-0.642
Second set	19 (324)	0.509	0.641	-0.003	0.115	0.521	0.240-0.698
κ Values within centers[a]							
Reggio Emilia	6 (186)	0.770	0.770	0.486	0.822	0.7425	0.591-0.864
Florence	2 (84)	1	1	1	1	1	
Turin	5 (177)	0.786	0.836	0.239	0.766	0.770	0.671-0.882
Perugia	3 (126)	0.967	1	-	0.742	0.967	0.908-1.000

Abbreviation: CI, confidence interval.
[a] This analysis included only centers in which more than 1 reader evaluated both sets of slides.

0.616) for nonexperienced centers. The κ values were similar for the first and second sets (overall: $\kappa = 0.505$ [95% CI, 0.358-0.642] and $\kappa = 0.521$ [95% CI, 0.240-0.698], respectively) (Table 2).

The 10 slides (slides 1.5, 1.7, 1.8, 2.6, 2.10, 3.2, 3.8, 4.1, 4.3, and 4.5) that obtained particularly low agreement in the first interpretation (the original κ value among all readers for these slides was 0.243; 95% CI, 0.139-0.359)

TABLE 3. Intraobserver Agreement for the 10 Slides With the Most Discordant Results That Were Re-Evaluated by 6 Experienced Readers

	Second Evaluation				
First Evaluation	Negative	Positive	Inconclusive	Inadequate	Total
Negative	20	1	1	0	22
Positive	4	23	0	0	27
Inconclusive	2	0	1	0	3
Inadequate	0	0	0	1	1
Total	26	24	2	1	53

were chosen for reinterpretation by 6 experienced readers. The results from the first and second evaluations are provided in Table 3. The κ value among the 6 expert readers from the interpretation of these 10 slides in the first reading was 0.616 (95% CI, 0.384-0.866) and decreased slightly in the second reading to 0.403 (95% CI, 0.182-0.643). Intraobserver agreement in this subset was 0.732 (95% CI, 0.694-0.751).

DISCUSSION

The lack of common criteria and of a consensus threshold for the evaluation of $p16^{ink4a}$ positivity in cytologic samples has generated great variability in the interpretation of this biomarker, partly because of the controversy regarding the simultaneous use of morphologic evaluation, which often is subjective and is especially difficult in cases of equivocal abnormalities. According to dual-staining criteria, positivity is scored independent of cell morphology. This aspect is also essential, because an ideal biomarker should be as independent as possible by subjective interpretation.[26] In the current study, we observed good reproducibility in terms of the interpretation of dual-staining for $p16^{ink4a}$ and Ki-67. The reproducibility was slightly higher when only the reports from skilled readers were considered, but it did not significantly increase from the first to the second set of slides after a second meeting in which the slides with the most discordant interpretations were discussed. Unsurprisingly, the agreement was very good for the positive and negative categories, whereas it was null for the inconclusive category. Conversely, the second reading of the same slides by expert readers unexpectedly did not improve on agreement and even decreased it (from 0.616 to 0.403). However, it should be noted that, for some slides, the second interpretation was performed more than 2 years after immunostaining, which caused the staining to fade and made evaluations more difficult. In other cases, the interpretation of discordant slides was especially ambiguous and challenging, even for experienced readers. This is consistent with our observation that the intraobserver agreement in the second interpretation was similar to the inter-observer agreement found in the first interpretation.

The results presented in this study are consistent with those from the 2 previous studies that assessed dual-staining reproducibility,[19,20] although the κ value in our study was slightly lower (0.61 vs nearly 0.7 reported by Allia et al[19] and 0.65-0.81 reported by Wentzensen et al[20]). This difference can be explained by our evaluation system, which used 4 categories, including the inconclusive category, compared with the 2-category system used in the other studies. Moreover, although the number of slides in this study was lower than that in the other 2 articles, to the best of our knowledge, this is the first study involving a very large number of participants. Indeed, no more than 2 laboratories participated in the previous studies. Our findings clearly demonstrated that the agreement within laboratories was much higher than the overall agreement.

All 3 of the studies measuring the reproducibility of the dual staining, including this report, considered samples from HPV-positive women, because, currently, in screening based on HPV DNA as the primary test, triage for HPV-positive women is the most plausible use of the $p16^{ink4a}$/Ki-67 dual-staining test. To date, the prevalent triage test in HPV-based screening programs is cervical cytology. Unfortunately, most of the studies assessing the inter-reader agreement for cervical cytology have been conducted in populations with unknown HPV status, although a recent report demonstrated that the accuracy of cervical cytology can be further improved by knowledge of the HPV status.[27] Although the studies on cytology have some methodological differences in terms of the populations and the number of categories used for cytology evaluation compared with $p16^{ink4a}$/Ki-67 evaluation, the κ

values obtained in our study were only slightly better than those obtained for cytology, which ranged from 0.56 to 0.60, in similar studies.[28,29] Because the experience using biomarkers is still limited compared with the experience using morphology alone, our results suggest that the efforts required to obtain good agreement with p16ink4a/Ki-67 dual staining are probably much less than those required for cytology. In fact, the good agreement in cervical cytology has been obtained after decades of use in large laboratories and thanks to continuous work by the scientific and clinical community.[30,31] Nevertheless, the high agreement reached in cervical cytology in countries with well implemented screening programs suggests that, in the short term, we cannot expect a strong increase in the reproducibility of p16ink4a/Ki-67 dual staining used as the triage test instead of cytology.

The dual-staining assay has been introduced to decrease the need for morphologic evaluation of p16ink4a-stained cells and thus to reduce the subjectivity of the evaluation. Although there is no doubt that the reproducibility of dual-staining interpretation is good to excellent, it is difficult to establish the extent to which adding Ki-67 staining improves the test reproducibility. In fact, a study evaluating the reproducibility of p16ink4a immunostaining alone on cytologic specimens demonstrated very high agreement and κ values, although it took place in an experimental setting and with a limited number of different readers.[32] In another study, however, dual staining with p16ink4a/Ki-67 demonstrated better specificity compared with p16 alone.[33]

Limits

In our study we primarily aimed at generating agreement on the criteria to be used in the NTCC2 trial; and, as such, we decided to classify ambiguous cases as inconclusive and to discuss them in a further plenary session. This classification, by definition, decreases the overall κ value and is obviously not acceptable in clinical practice. In a clinical setting, these cases would give rise to test repetition or inadequate reports. Furthermore, the first set of slides was not drawn at random, but it was selected to include a large number of ambiguous cases; therefore, it did not represent a real population of HPV-positive women. This surely led to an almost inevitable reduction in the κ values that we would have expected if the population had included only randomly chosen HPV-positive women. Therefore, it is likely that, in real practice, the reproducibility of the p16ink4a/Ki-67 dual-staining test would be higher.

Conclusions

The interpretation of dual staining for p16ink4a and Ki-67 demonstrated good reproducibility when evaluated in a large number of laboratories, confirming its robustness, which is an essential prerequisite for its consideration as a triage test for HPV-positive women in cervical cancer screening programs that adopt HPV as the primary test. Further investigations are needed to evaluate the accuracy and clinical utility of the test.

FUNDING SUPPORT

This study is part of the New Technologies for Cervical Cancer Screening 2 (NTCC2) trial funded by the Italian Ministry of Health, which owns the data (grant RF-2009-1536040).

CONFLICT OF INTEREST DISCLOSURES

Maria Benevolo reports nonfinancial support from Roche Diagnostics and Hologic outside the submitted work for a study funded by the Italian Ministry of Health, data owner. Paolo Giorgi Rossi reports nonfinancial support from Roche Diagnostics and Hologic Genprobe outside the submitted work for a study funded by the Italian Ministry of Health, data owner. The remaining authors made no disclosures.

AUTHOR CONTRIBUTIONS:

Maria Benevolo: Conceptualization, methodology, investigation, resources, data curation, writing–original draft, supervision, and project administration. **Elena Allia**: Methodology, validation, investigation, and writing–review and editing. **Daniela Gustinucci**: Methodology, investigation, resources, and project administration. **Francesca Rollo**: Methodology, investigation, resources, and writing–review and editing. **Simonetta Bulletti**: Methodology, investigation, and resources. **Elena Cesarini**: Methodology, investigation, and resources. **Basilio Passamonti**: Investigation, resources, and project administration. **Maria Rosaria Giovagnoli**: Investigation, resources, and writing–review and editing. **Elisabetta Carico**: Methodology, investigation, and writing–review and editing. **Francesca Carozzi**: Conceptualization, validation, investigation, writing–original draft, and writing–review and editing.

Alessandra Mongia: Investigation and resources. **Giulia Fantacci**: Investigation and resources. **Massimo Confortini**: Methodology, investigation, and writing–review and editing. **Teresa Rubino**: Validation. **Cristina Fodero**: Validation. **Sonia Prandi**: Methodology, validation, investigation, and writing–review and editing. **Natalina Marchi**: Methodology, validation, investigation, and writing–review and editing. **Angelo Farruggio**: Methodology, validation, investigation, and writing–review and editing. **Anna Coccia**: Investigation. **Luigia Macrì**: Methodology, investigation, and writing–review and editing. **Bruno Ghiringhello**: Methodology, validation, investigation, and writing–review and editing. **Guglielmo Ronco**: Conceptualization. **Emma Bragantini**: Validation, investigation, and writing–review and editing. **Enzo Polla**: Investigation, resources, and data curation. **Vincenzo Maccallini**: Methodology, validation, investigation, resources, and data curation. **Giovanni Negri**: Methodology, investigation, and writing–original draft. **Paolo Giorgi Rossi**: Conceptualization, methodology, formal analysis, resources, data curation, writing–original draft, writing–review and editing, supervision, funding acquisition.

REFERENCES

1. Ronco G, Dillner J, Elfstrom KM, et al. Efficacy of HPV-based Screening for Preventing Invasive Cervical Cancer: follow-up of European randomised controlled trials. *Lancet*. 2014;383:524-532.
2. Anttila A, Arbyn M, de Vuyst H, et al. eds. European Commission European Guidelines for Quality Assurance in Cervical Cancer Screening. 2nd ed. Supplements. Luxembourg: Office for Official Publications of the European Communities; 2015.
3. Saslow D, Solomon D, Lawson HW, et al. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *Am J Clin Pathol*. 2012;137:516-542.
4. Moyer VA, US Preventive Services Task Force. Screening for cervical cancer: US Preventive Services Task Force recommendation statement. *Ann Intern Med.* 2012;156:880-891.
5. von Knebel Doeberitz M. New markers for cervical dysplasia to visualise the genomic chaos created by aberrant oncogenic papillomavirus infections. *Eur J Cancer*. 2002;38:2229-2242.
6. Carozzi F, Confortini M, Dalla Palma P, et al. and the NTCC Working Group. Use of p16 overexpression to increase the specificity of human papillomavirus testing: a study nested in the NTCC randomised controlled trial. *Lancet Oncol*. 2008;9:937-945.
7. Bergeron C, Ikenberg H, Sideri M, et al. and PALMS Study Group. Prospective evaluation of p16/Ki-67 dual-stained cytology for managing women with abnormal Papanicolaou cytology: PALMS study results. *Cancer Cytopathol*. 2015;123:373-3781.
8. Carozzi F, Gillio-Tos A, Confortini M, et al. Risk of high-grade cervical intraepithelial neoplasia during follow-up in HPV-positive women according to baseline p16-INK4A results: a prospective analysis of a nested substudy of the NTCC randomized controlled trial. *Lancet Oncol*. 2013;14:168-176.
9. Gustinucci D, Giorgi Rossi P, Cesarini E, et al. Use of cytology, E6/E7 mRNA and p16INK4a-Ki67 to define the management of human papillomavirus (HPV) positive women in cervical cancer screening. *Am J Clin Pathol*. 2016;145:35-45.
10. Fehrmann F, Laimins LA. Human papillomaviruses: targeting differentiating epithelial cells for malignant transformation. *Oncogene*. 2003;22:5201-5207.
11. zur Hausen H. Immortalization of human cells and their malignant conversion by high risk human papillomavirus genotypes. *Semin Cancer Biol*. 1999;9:405-411.
12. Dona MG, Vocaturo A, Giuliani M, et al. p16/Ki-67 dual staining in cervico-vaginal cytology: correlation with histology, human papillomavirus detection and genotyping in women undergoing colposcopy. *Gynecol Oncol*. 2012;126:198-202.
13. Tornesello ML, Buonaguro L, Giorgi-Rossi P, Buonaguro FM. Viral and cellular biomarkers in the diagnosis of cervical intraepithelial neoplasia and cancer [serial online]. *Biomed Res Int* 2014: 519619, 2013.
14. Liang J, Mittal KR, Wei JJ, Yee H, Chiriboga L, Shukla P. 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; 26:71-75.
15. Longatto Filho A, Utagawa ML, Shirata NK, et al. Immunocytochemical expression of p16INK4A and Ki-67 in cytologically negative and equivocal pap smears positive for oncogenic human papillomavirus. *Int J Gynecol Pathol*. 2005;24:118-124.
16. Aoyama C, Liu P, Ostrzega N, Holschneider CH. Histologic and immunohistochemical characteristics of neoplastic and nonneoplastic subgroups of atypical squamous lesions of the uterine cervix. *Am J Clin Pathol*. 2005;123:699-706.
17. Tsoumpou I, Arbyn M, Kyrgiou M, et al. p16INK4a immunostaining in cytological and histological specimens from the uterine cervix: a systematic review and meta-analysis. *Cancer Treat Rev*. 2009;35:210-220.
18. Wentzensen N, Bergeron C, Cas F, Eschenbach D, Vinokurova S, von Knebel DM. Evaluation of a nuclear score for p16INK4a-stained cervical squamous cells in liquid-based cytology samples. *Cancer*. 2005;105:461-467.
19. Allia E, Ronco G, Coccia A, et al. Interpretation of p16(INK4a)/Ki-67 dual immunostaining for the triage of human papillomavirus-positive women by experts and nonexperts in cervical cytology. *Cancer Cytopathol*. 2015;123:212-218.
20. Wentzensen N, Fetterman B, Tokugawa D, et al. Interobserver reproducibility and accuracy of p16/Ki-67 dual-stain cytology in cervical cancer screening. *Cancer Cytopathol*. 2014;122:914-920.
21. Solomon D, Davey D, Kurman R, et al. Bethesda 2001 Workshop. The 2001 Bethesda System: terminology for reporting results of cervical cytology. *JAMA*. 2002;287:2114-2119.
22. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76:378-382.
23. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20:37-46.
24. Reichenheim ME. Confidence intervals for the kappa statistic. *Stata J*. 2004;4:421-428.
25. Stata Corporation. Stata Statistical Software. Release 13 [software program]. College Station, TX: StataCorp LP; 2013.
26. Arbyn M, Ronco G, Cuzick J, Wentzensen N, Castle PE. How to evaluate emerging technologies in cervical cancer screening? *Int J Cancer*. 2009;125:2489-2496.
27. Bergeron C, Giorgi-Rossi P, Cas F, et al. Cytology for triaging HPV-positive women: substudy nested in the NTCC randomized controlled trial [serial online]. *J Natl Cancer Inst*. 2015;107. pii: dju423.

28. Stoler MH, Schiffman M; Atypical Squamous Cells of Undetermined Significance-Low-Grade Squamous Intraepithelial Lesion Triage Study (ALTS) Group. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL Triage Study. *JAMA.* 2001;285:1500-1505.
29. Confortini M, Bondi A, Cariaggi MP, et al. Interlaboratory reproducibility of liquid-based equivocal cervical cytology within a randomized controlled trial framework. *Diagn Cytopathol*. 2007;35:541-544.
30. Confortini M, Di Stefano C, Biggeri A, et al. Daily peer review of abnormal cervical smears in the assessment of individual practice as an additional method of internal quality control. *Cytopathology*. 2016;27:35-42.
31. Branca M, Alieri S, Cialdea L, Morosini P. Survey of performance of cervical cytopathology laboratories and of screening programs in Italy. National Working Group for Quality Assurance in Cytohistopathology. *Tumori*. 1990;76:434-438.
32. Vinyuvat S, Karalak A, Suthipintawong C, et al. Interobserver reproducibility in determining p16 overexpression in cervical lesions: use of a combined scoring method. *Asian Pac J Cancer Prev*. 2008;9653-9657.
33. Schmidt D, Bergeron C, Denton KJ, Ridder R; European CINtec Cytology Study Group. p16/ki-67 dual-stain cytology in the triage of ASCUS and LSIL Papanicolaou cytology: results from the European equivocal or mildly abnormal Papanicolaou cytology study. *Cancer Cytopathol.* 2011;119:158-166.