research article

Evaluation of the training program for p16/Ki-67 dual immunocytochemical staining interpretation for laboratory staff without experience in cervical cytology and immunocytochemistry

Veronika Kloboves Prevodnik[1,2], Ziva Pohar Marinsek[1], Janja Zalar[1], Hermina Rozina[1], Nika Kotnik[3], Tine Jerman[4], Jerneja Varl[2,5], Urska Ivanus[4,2]

[1] Department of Cytopathology, Institute of Oncology Ljubljana, Ljubljana, Slovenia
[2] Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
[3] Department of Experimental Allergology and Immunodermatology Oldenburg, Carl von Ossietzky Universität, Oldenburg, Germany
[4] Epidemiology and Cancer Registry, Institute of Oncology Ljubljana, Ljubljana, Slovenia
[5] Department of Experimental Oncology, Institute of Oncology Ljubljana, Ljubljana, Slovenia

Radiol Oncol 2020; 54(2): 201-208.

Received 22 January 2020
Accepted 4 March 2020

Correspondence to: Assoc. Prof. Veronika Kloboves Prevodnik, M.D. Ph.D., Department of Cytopathology, Institute of Oncology Ljubljana, Zaloška 2, SI-1000 Ljubljana, Slovenia. E-mail: vkloboves@onko-i.si

Disclosure: No potential conflicts of interest were disclosed.

Background. p16/Ki-67 dual immunocytochemical staining (DS) is considered easy to interpret if evaluators are properly trained, however, there is no consensus on what constitutes proper training. In the present study we evaluated a protocol for teaching DS evaluation on students inexperienced in cervical cytology.

Methods. Initial training on 40 DS conventional smears was provided by a senior cytotechnologist experienced in such evaluation. Afterwards, two students evaluated 118 cases. Additional training consisted mainly of discussing discrepant cases from the first evaluation and was followed by evaluation of new 383 cases. Agreement and accuracy of students' results were compared among the participants and to the results of the reference after both evaluations. We also noted time needed for evaluation of one slide as well as intra-observer variability of the teacher's results.

Results. At the end of the study, agreement between students and reference was higher compared to those after initial training (overall percent agreement [OPA] 81.4% for each student, kappa 0.512 and 0.527 vs. OPA 78.3% and 87.2%, kappa 0.556 and 0.713, respectively). However, accuracy results differed between the two students. After initial training sensitivity was 4.3% points and 2.9% points higher, respectively compared to the reference, while specificity was 30.6% points and 24.4% points lower, respectively, compared to the reference. At the end of the study, the sensitivity reached by one student was the same as that of the reference, while it was 2.6% points lower for the other student. There was a statistically significant difference in specificity between one student and the reference and also between students (16.7 and 15.1% points). Towards the end of the study, one student needed 5.2 min for evaluating one slide while the other needed 8.2 min. The intra-observer variability of the senior cytotechnologist was in the range of "very good" in both arms of the study.

Conclusions. In teaching DS evaluation, the students' progress has to be monitored using several criteria like agreement, accuracy and time needed for evaluating one slide. The monitoring process has to continue for a while after students reach satisfactory results in order to assure a continuous good performance. Monitoring of teacher's performance is also advisable.

Key words: training protocol; p16/Ki-67 dual immunocytochemical staining; agreement; accuracy; inter-observer reproducibility

Introduction

p16/Ki-67 dual immunocytochemical staining (DS) is considered easy to interpret if evaluators are properly trained. However, there is no consensus on what constitutes proper training. Authors have used different training approaches in studies investigating inter-observer reproducibility and accuracy of DS.[1-6] Most training protocols described in these studies consisted of initial and additional training. The initial training was provided by the manufacturer, however, it was not exactly the same in all cases except that it was completed by a proficiency test. The information on the initial training is sparse. Three of the above mentioned studies do not describe the initial training.[1,2,4] In two studies, participants were shown 15[5] or 40[6] microscope-projected images, while in the third study participants examined a teaching set of slides.[3] The number of cases in the teaching set is not mentioned. In all three studies the training was completed in one or two-half day sessions.[3,5,6] Four authors described additional training which consisted of evaluating from 80 to 469 slides as well as reviewing and discussing cases with discrepant results.[1-3,6] Agreement in DS interpretation among evaluators improved after additional training in all three studies (kappa range: 0.43–0.73 compared to 0.50–0.87).[1-3,6] In the study of Wentzensen *et al*.[2] agreement was evaluated only at the end of the study.

In our recently published study which assessed reproducibility of the DS test we described a training protocol which was designed to introduce DS in three Slovenian cytopathological laboratories participating in the national organized cervical cancer screening program.[6] At the time we designed the protocol we found only one similar study by Waldstrom *et al*.[1] The results of our study demonstrated that initial training by the manufacturer was not enough for achieving accurate results of DS interpretation. Furthermore, the manufacturer provides training when an institution is ready to implement the test. Later on, when we have the need to teach additional personnel, we have to have our own training protocol which will assure the students receive the necessary expertise.

On the basis of the results of our previous study, we proposed a training protocol for staff inexperienced in DS reading.[6] In the present study, we aimed to test the proposed training protocol for DS interpretation on two students inexperienced in DS reading and to discern how improvement in DS evaluation influences the time needed for DS reading of one slide. An additional end point of the study was also monitoring the performance of the senior cytotechnologist involved in teaching DS interpretation to new personnel.

Material and methods

Study design and setting

We used DS slides on conventional cervical smears taken from 501 women who underwent colposcopy at Celje General Hospital or at University Medical Centre Maribor between April 2014 and December 2015. Samples from 118 women were the same ones we have used in our previous study.[6] These women were invited to colposcopy per screening program guidelines, either due to high-grade (HG) cytology, a human Papillomavirus (HPV)-positive triage test after low-grade pathological changes or due to a positive HPV test during follow-up after treatment of high grade cervical intraepithelial neoplasia (CIN). An additional set of samples came from 383 women, of which 87 were referred to colposcopy from the screening program and 296 were non-responders. All non-responders were invited to colposcopy after they have taken their own cervical sample. 250 were HPV positive and 46 were HPV negative.[7] Sample acquisition, procedures following abnormal colposcopy, reasons for excluding patients from the study and classification of histopathology results were the same as already described in our previous study.[6] In both sets of women the same gynecologists participated in colposcopy examinations and the same criteria were used for performing biopsy for histological examinations. After the initial colposcopy, all women were followed via the Cervical Cancer Screening Registry ZORA that registers all cervical cytology, HPV test results and cervical histology results of all Slovenian women.

Two students, a biologist (S1) and a medical doctor (S2) as well as a senior cytotechnologist (SC) participated in the study. SC was employed at the Department of cytopathology, Institute of Oncology Ljubljana, where the study was conducted. She was trained in DS evaluation during our previous study. The two students had no previous knowledge of cervical cytology or of DS evaluation. Four cytopathologist from the Institute of Oncology Ljubljana reviewed all DS slides. Their consensus results were considered as reference.

Fixation of slides, immunocytochemical staining and the rules for slide interpretation have already been described in our previous paper.[6]

The reading of DS slides was divided into primary reading of 118 slides after initial training and the secondary reading of new 383 slides after additional training. In each reading, students spotted the DS cells and passed slides on to the SC who reviewed all of them. The SC reviewed both sets of slides separately after each student and therefore, each case had four readings. Results of both students and of the SC were evaluated after initial and after secondary reading and compared to the reference results. For both students we also noted the time needed to interpret each of the 501 slides. S1 evaluated 501 slides during the course of four months while S2 evaluated all the slides in five months because none of the students were evaluating slides continuously eight hours per day.

The p16/Ki-67 DS study was nested within the randomized trial of HPV self-sampling among non-attenders of ZORA. It was conducted in compliance with the Helsinki Declaration, and was approved by the institutional review board and the National Medical Ethics Committee at the Slovenian Ministry of Health (consents Nos. 155/03/13 and 136/04/14). All women signed informed consent to participate in the study. This research was financed by the Slovenian Research Agency and the Slovenian Ministry of Health (trial No. L3-5512).

Training design

The initial training program for DS interpretation started by lectures and by demonstration of morphology of normal and atypical cervical cytology and of DS interpretation. Afterwards, the students examined 40 teaching slides and discussed difficult cases with SC at a multi-head microscope. Training was completed in one week. Additional training took place after we evaluated the results of the primary reading. It was also provided by SC and lasted two days. Additional training included a troubleshooting slide review of discordant cases of the primary reading as well as a theoretical repetition of the criteria for DS evaluation.

Study outcomes and statistical analysis

The primary outcomes were: (1) agreement in DS interpretation between all three evaluators (S1, S2, SC), between each evaluator and the reference as well as intra-observer agreement for SC; (2) accuracy prior to and after the additional training. For evaluating agreement and accuracy we used the same statistical methods as already described in our previous article.[6] The secondary outcome was measuring screening time per slide of both students. We assessed the mean screening time with standard deviation (SD) for primary and secondary evaluation and also for the last 100 slides. The screening time trends for positive and negative p16/Ki-67 DS results were characterized in terms of an average percent change per slide estimated by the log-linear joinpoint regression. Sensitivity, specificity, positive (PPV) and negative (NPV) predictive values were calculated for one year histopathological follow-up for both evaluation sets. In addition, sensitivity and specificity were calculated at each of the 501 ratings, taking into account only 100 most recent ratings, and presented on a line plot.

We conducted all our analyses with R v3.5.1[8], using 2-tailed tests and the significance level α = 0.050. Joinpoint Regression program[9] was used for the assessment of the screening time trends.

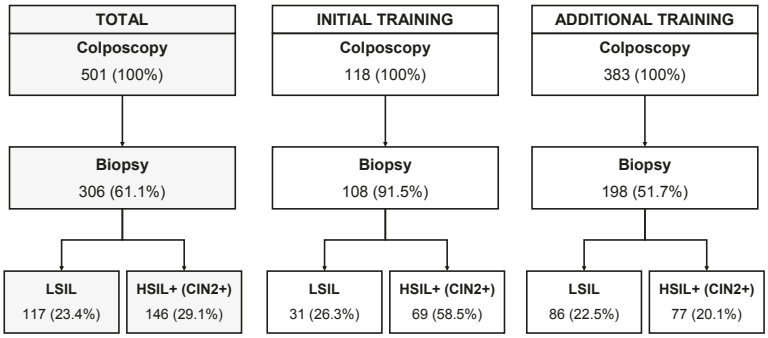

FIGURE 1. Study flow chart with histopathology follow-up results.

Results

Study population

The average age of the 501 women in the study was 44.3 years (SD 11.9). The average age of women whose smears were subject to primary DS evaluation was 36.5 years (SD 11.1) while the average age of women in secondary DS evaluation was 46.7 years (SD 11.2). The study flow chart with histopathology results is shown in Figure 1. Women whose smears were included into the primary DS evaluation had higher prevalence of cervical intraepithelial neoplasia grade 2 or worse (CIN2+) compared to women whose samples were material for the secondary evaluation (58.5 % *vs.* 20.1%) (Figure 1).

TABLE 1. p16/Ki-67 study results and CIN2+ outcome for students, senior cytotechnologist and reference

Reviewer	Categories of p16/Ki67 dual staining result	Initial training (N = 118)		Additional training (N = 383)	
		p16/Ki67 dual staining result N, (%)	CIN2+ outcomes N (PV, %)*	p16/Ki67 dual staining result (N, %)	CIN2+ outcomes N (PV, %)*
S1	positive	91 (77.1)	65 (71.4)	171 (44.6)	65 (38.0)
	suspicious	7 (5.9)	2 (28.6)	7 (1.8)	2 (28.6)
	negative	20 (16.9)	2 (10.0)	205 (53.5)	10 (4.9)
	unsatisfactory	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)
S2	positive	92 (78.0)	66 (71.7)	129 (33.7)	64 (49.6)
	suspicious	2 (1.7)	0 (0.0)	1 (0.3)	1 (100.0)
	negative	24 (20.3)	3 (12.5)	253 (66.1)	12 (4.7)
	unsatisfactory	0 (0)	0 (0.0)	0 (0.0)	0 (0.0)
SC (S1)	positive	83 (70.3)	65 (78.3)	143 (37.9)	65 (45.5)
	suspicious	2 (1.7)	2 (100.0)	5 (1.3)	0 (0.0)
	negative	32 (27.1)	2 (6.3)	235 (61.4)	12 (5.1)
	unsatisfactory	1 (0.8)	0 (0.0)	0 (0.0)	0 (0.0)
SC (S2)	positive	83 (70.3)	65 (78.3)	145 (37.9)	64 (44.1)
	suspicious	2 (1.7)	2 (100.0)	4 (1.0)	0 (0.0)
	negative	32 (27.1)	2 (6.3)	234 (61.1)	13 (5.6)
	unsatisfactory	1 (0.8)	0 (0.0)	0 (0.0)	0 (0.0)
Reference	positive	78 (66.1)	64 (82.1)	114 (29.8)	64 (56.1)
	suspicious	2 (1.7)	0 (0)	13 (3.4)	3 (23.1)
	negative	38 (32.2)	5 (13.2)	255 (66.6)	10 (3.9)
	unsatisfactory	0 (0.0)	0 (0.0)	1 (0.3)	0 (0.0)

N = number of cases; Reference = results of four cytopathologists at the Deptment of Cytopathology, Institute of Oncology Ljubljana; S1 = student 1; S2 = student 2; SC (S1) = senior cytotechnologist results obtained during revision of student 1 results; SC (S2) = senior cytotechnologist results obtained during revision of student 2 results; * PV = predictive value (number of CIN2+detected within specific category of p16/Ki-67 dual staining result divided by the number of test results in specific category)

TABLE 2. Individual comparison of p16/Ki-67 agreement and performance between students, senior technologist and reference

Training	Reviewers	OPA	McNemar's test p	κ (cohen) (95% CI)
Initial training (N = 118)	S1/Reference	81.4%	0.000	0.512 (CI: 0.329–0,696)
	S2/Reference	81.4%	0.006	0.527 (CI: 0.349–0,705)
	SC (S1)/ Reference	94.1%	0.131	0.859 (CI: 0.758–0.960)
	SC (S2)/ Reference	94.1%	0.131	0.859 (CI: 0.758–0.960)
	SC (S1)/SC (S2)	100.0%	/	1.000 (CI: /)
Addtional (N = 383)	S1/Reference	78.3%	0.000	0.556 (CI: 0.472–0.641)
	S2/Reference	87.2%	0.775	0.713 (CI: 0.638–0.788)
	SC (S1)/ Reference	86.2%	0.006	0.700 (CI: 0.625–0,775)
	SC (S2)/ Reference	86.4%	0.004	0.707 (CI: 0.632–0.781)
	SC (S1)/SC (S2)	98.7%	1.000	0.973 (CI: 0.949–0.996)

N = number of cases; OPA = overall percent agreement; S1 = student 1; S2 = student 2; SC (S1) = senior cytotechnologist results obtained during revision of student 1 results; SC (S2) = senior cytotechnologist results obtained during revision of student 2 results; SC1/SC2 = intra-observer variability; Reference = results of four cytopathologists at the Deptment of Cytopathology, Institute of Oncology Ljubljana; * Scale for interpretation of κ values = below 0.20 (poor), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (good), >0.81 (very good)[10]

p16/Ki-67 dual staining results and agreement

In primary and secondary evaluation both students (S1 and S2) had more positive p16/Ki-67 DS results compared to the reference (Table 1). However, in secondary evaluation, S2 had less positive results (33.7%) compared to the outcome in primary evaluation. Her percentage of positive results was even closer to the reference results (29.8%) than that of SC (37.9%). All evaluators used the suspicious category sparsely, only 0.3–5.9% of results fell into this category.

In primary evaluation, the agreement of DS results between reference and each of the students was moderate, while the agreement between reference and SC was very good. However, S2 reached good agreement in secondary evaluation which even slightly surpassed the agreement result between SC and the reference (Table 2). The agreement between S1 and the reference remained mod-

erate. Intra-observer variability between the two evaluations performed by the SC after each student was very good in primary, as well as in secondary evaluation (Table 2).

Accuracy of p16/Ki-67 results

Reference results showed higher sensitivity and positive predictive value for CIN2+ in primary compared to secondary evaluation (92.8% *vs.* 87.0% and 80.0% *vs.* 52.8%, respectively) (Supplementary Table 1, Figure 2). Specificity and negative predictive value for CIN2+ were lower in primary compared to secondary evaluation (67.3% *vs.* 80.4% and 86.8% *vs.* 96.1%, respectively).

In primary evaluation, the results of both students had slightly higher sensitivity and negative predictive values for CIN2+ but much lower specificity and positive predictive values compared to reference results (Supplementary Table 1, Figure 2, Figure 3). While students preformed similarly in terms of sensitivity and specificity after initial training, after additional training the results of S2 were closer to the results of the reference compared to the results of S1. In secondary evaluation, S2 even reached higher sensitivity and specificity than SC (84.4% *vs.* 83.1% and 78.8% *vs.* 72.2%, respectively). S1 had the highest sensitivity in both evaluations (97.1% and 87.0%) combined with the lowest specificity (36.7% and 63.7%, respectively) among all evaluators. The specificity of S1 was significantly lower than the reference's in both primary and secondary evaluation. Her specificity came close to the specificity of the reference after evaluating approximately 250 slides. However, after this point her performance started to decline (Figure 3).

Screening time

S1 evaluated 501 slides in 96 hours and 40 minutes while S2 needed 95 hours and 56 minutes for evaluating all the slides. In primary evaluation S1 needed less screening time

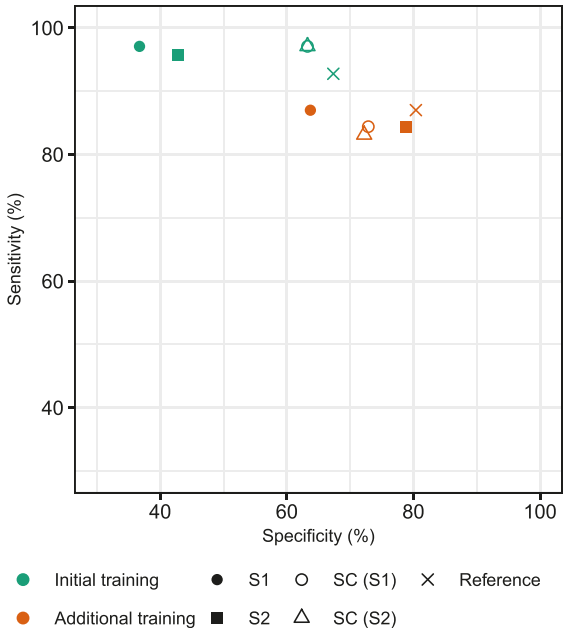

FIGURE 2. Sensitivity and specificity of p16/Ki-67 dual immunocytochemical staining (DS) for detecting CIN2+ for both students and the teacher (senior cytotechnologist).

S1= student 1; S2 = student 2; SC (S1) = senior cytotechnologist – slide review after S1; SC (S2) = senior cytotechnologist – slide review after S2

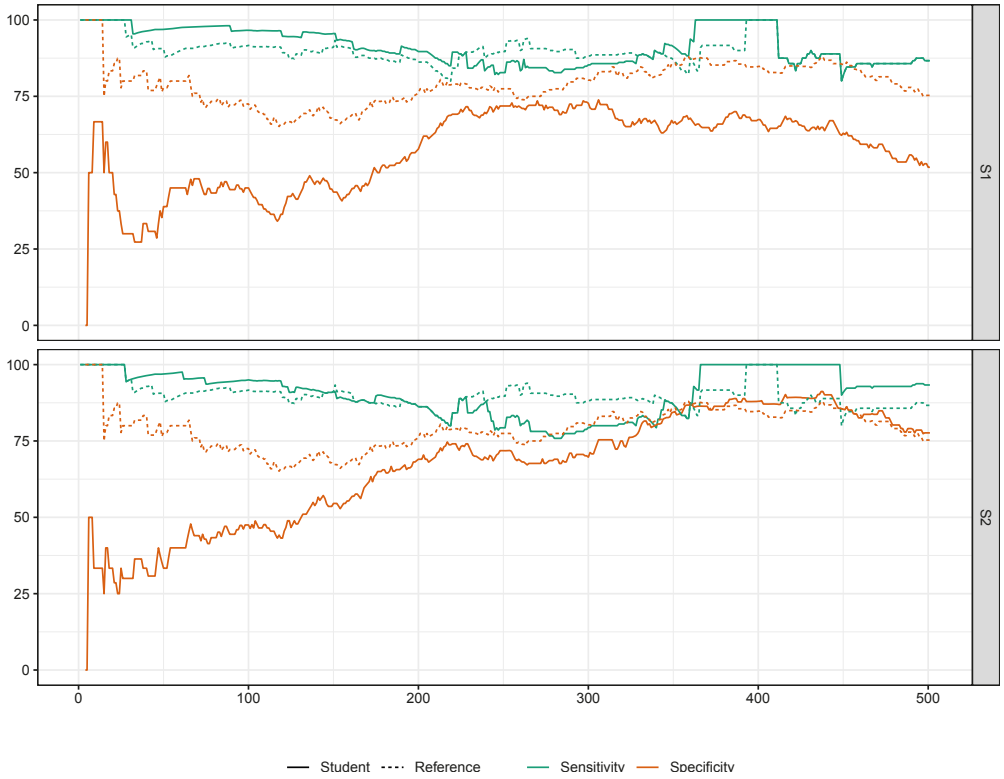

FIGURE 3. Sensitivity and specificity for the results of both students and the reference according to the number of evaluated slides.

S1 = student 1; S2 = student 2

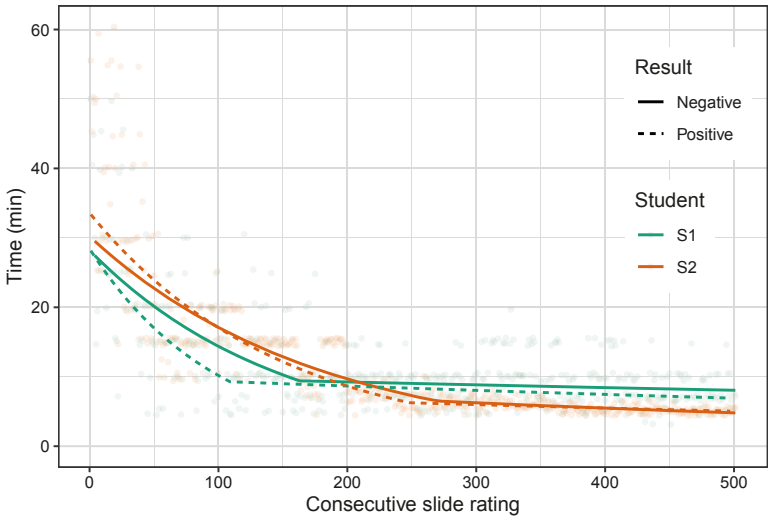

FIGURE 4. Joinpoint regression analysis of students' screening times.

S1 = student 1; S2 = student 2

per slide (18.8 min; SD 10.3) compared to S2 (24.2 min; SD 13.0; $p < 0.001$). In secondary evaluation, the screening times decreased for both students, however, S1 now needed more time (9.3 min; SD 3.9) compared to S2 (7.6 min; SD 3.5; $p < 0.001$). At the end of the training, during the evaluation of the last 100 slides, the average time of two students was 6.7 minutes per slide. However, the difference between them was statistically significant (S1: 8.2 min *vs.* S2: 5.2 min; SD 2.8 and 0.4, respectively; $p < 0.001$).

Joinpoint regression analysis of screening times of S1 showed the joinpoint at 163rd slide for negative slides. There was a 0.7% decrease per each slide ($p < 0.05$) in the first segment and 0.0% decrease in the second segment. The joinpoint for positive slides was at 109th slide with 1.0% decrease per slide ($p < 0.05$) in the first segment and 0.1% decrease ($p < 0.05$) in the second segment (Figure 4). For S2, the joinpoint for negative slides was at 207th slide with 0.6% decrease per slide ($p < 0.05$) in the first segment and 0.1% decrease ($p < 0.05$) in the second segment. The joinpoint for positive slides in case of S2 was at 248th slide with 0.6% decrease ($p < 0.05$) in the first segment and 0.1% decrease ($p < 0.05$) in the second segment.

Discussion

The results of our study confirmed that the training protocol we have used was adequate for teaching the interpretation of p16/Ki-67 DS. At the end of the training one student was competent for independent DS interpretation as evidenced by comparable accuracy results and by good agreement between her results and those of the reference. Towards the end of the training this student needed 5.2 min for evaluating one slide. The training protocol is suitable also for monitoring the performance of the teacher.

We do not have an explanation as to why the results of the other student showed a statistically significant difference in agreement and accuracy when compared to the results of the reference. Since environmental factors were the same for both students, it seems that internal factors played an important role in performance difference. It is well known that interpreting slides is partly a subjective method. Furthermore, in our previous study we demonstrated a number of reasons that contributed significantly to the difficulty of DS interpretation.[6] Weak p16 staining, less preserved cell morphology and strong background staining were important drawbacks of decision making. Similar observations were also made by McMenamin *et al.*[4] and Benevolo *et al.*[5] An additional important information from our previous study was also the fact that 45% of cases which were marked as suspicious for p16/Ki-67 positivity had only one such cell.[6] Therefore, this type of training is not suitable for every person without prior knowledge of DS evaluation. However, by monitoring student's results, we can assess if additional experience to reach the necessary expertise is needed.

Most articles which describe agreement and accuracy of DS interpretation among observers briefly mention the training program they have used. Only the training programs in three studies had some similarities with our own.[1,2,3] However, the initial training provided by the manufacturer is not described in detail in any of them. The secondary training can be compared to ours because it consisted of reviewing a certain number of slides (80, 150 and 469) and of discussing discrepant cases of the first viewing. In the second reading, Allia *et al.*[3] mention evaluation of 350 new slides, while Wentzesen *et al.*[2] used 480 new slides, however, not all reviewers red all of them. In the study of Waldstrom *et al.*[1] they randomly selected 185 slides from set of 469 slides from the first reading. Only the study of Allia *et al.*[3] included three evaluators inexperienced in cervical cytology (two medical students and one biologist) in addition to the four experienced ones.

Unfortunately, we can compare our results to those of Allia *et al.*[3] only to a limited extent due to

differences in methodology. Allia *et al.* compared agreement between results of four evaluators with experience in DS evaluation and between results of three evaluators without experience. Agreement improved after additional training in each group. In our study, the population of 118 women from the first arm of the study had a higher percentage of histologically confirmed CIN2+ compared to the population of 383 women in the second arm (58.5% *vs.* 20.1%). This difference in the prevalence of the disease has to be taken into account when comparing the results of primary and secondary evaluation, especially Kappa values and predictive values of a test. In the Allia's *et al.* paper, the study population contained 14.7% of women with histological diagnoses of CIN2+.[3] Since the populations of women in the two studies were not totally alike, we can compare only the end results indirectly. At the end of Allia's *et al*. study, the specificity of DS for CIN2+ was 66.7% for the experienced evaluators, while the students reached a specificity of 60.5%, a difference of 5.2% points. On the basis of these data Allia *et al*. concluded that DS evaluation "can be performed even by staff not trained in the morphological interpretation of cytology" after a short training phase.[3] At the end of our study the difference in specificity between the reference and the successful student was 1.6% points and 16.7% points for the unsuccessful student. Therefore, we agree with Allia *et al.* that it is possible to perform DS evaluation with personnel not experienced in cervical cytology, however, not after a short training period.

If we translate 97 hours needed for reviewing 501 slides into 12 days with an eight-hour working day, the whole training period in our study would last 19 days. The evaluation of slides after initial training, as well as partly after additional training, has to be considered part of the learning process. This is clearly demonstrated in the graph of continual monitoring of the students' accuracy results. For the successful student the specificity continued rising during the evaluation of the first 118 slides. It continued to rise also for a period after the secondary training until approximately the time when the student evaluated roughly 350 slides altogether. At this point the accuracy results of the student were very similar to the results of the reference (same sensitivity, 87.0%) and slightly lower specificity (84.4 *vs.* 85.7%). The rest of the time used in the evaluation of the last 150 slides was necessary for monitoring the student's performance. The necessity of such action has proven to be correct in the case of the other student. Her accuracy results came close to those of the reference after she evaluated approximately 250 slides, however, her performance started to decline afterwards.

For monitoring the students' progress in DS evaluation it is advisable to use more than one criterion of successfulness. Using only agreement between the students and the reference will be reliable only when positive and negative predictive values of the reference results are high. However, these values depend partly on the percentage of CIN2+ cases within the population of women from which samples for DS evaluation originate. Therefore, comparing accuracy results between students and reference is beneficial because it will demonstrate more exactly where a particular student has difficulties in DS interpretation. For example, high sensitivity and low specificity point to the fact that a student is signing too many cases as positive. An additional measure of a student's successfulness is also the time needed to evaluate one slide.

Since both students were inexperienced in DS evaluation it is reasonable that the time needed to evaluate one slide was progressively decreasing. The decrease was sharper towards the beginning and less pronounced latter on. In addition to the agreement and the accuracy results, time needed for evaluating one slide also showed the difference between the two students. The ultimately unsuccessful student reached faster the point after which her time per slide started to decrease very slowly, compared to the successful student. However, during the evaluation of the last 100 slides, the successful student needed significantly less average time per slide compared to the other student (5.2 min *vs.* 8.2 min). Only McMenamin *et al.* made a quick mention of the time needed per slide evaluation.[4] They reported that their experienced DS evaluators needed less than 1 minute for evaluating a clearly positive slide and 3-4 min for more challenging ones. Their evaluation time per slide is shorter than in our study not only because our students were unexperienced but mainly because we used conventional smears while in the study of Mc Menamin *et al.* ThinPrep specimens were used.[4]

In addition to teaching DS interpretation to students, our training program was designed also to monitor the performance of the cytotechnologist involved in the teaching process. We believe that teacher monitoring is an important element of the training program which helps to assure the students will receive the best training. The intra-observer agreement was within the range of "very good" in both arms of our DS evaluation (OPA 100.0% and 98.7%, kappa 1.000 and 0.937, respec-

tively). This result is even slightly better than the results obtained by McMenamin *et al.*[4] where agreement for three cytotechnologists was 82.8% (0.65), 83.8% (0.67) and 94.9%, (0.91), respectively.

Conclusions

In conclusion we would like to say that teaching p16/Ki-67 interpretation should be a closely monitored process in which students' results have to be compared to the reference results with known accuracy for CIN2+. The students' progress has to be monitored using several criteria like agreement, accuracy and time needed for evaluating one slide. The monitoring process has to continue for a while after students reach satisfactory results in order to assure a continuous good performance. Monitoring of teacher's performance is also advisable.

References

1. Waldstrøm M, Christensen RK, Ørnskov D. Evaluation of p16 INK4a/Ki-67 dual stain in comparison with an mRNA human papillomavirus test on liquid-based cytology samples with low-grade squamous intraepithelial lesion. *Cancer Cytopathol* 2013; **121:** 136-45. doi: 10.1002/cncy.21233

2. Wentzensen N, Fetterman B, Tougawa D, Shiffman M, Castle PE, Wood SN, et al. Interobserver reproducibility and accuracy of p16/Ki67 dual-stain cytology in cervical cancer screening. *Cancer Cytopathol* 2014; **122:** 914-20. doi: 10.1002/cncy.21473

3. Allia E, Ronco G, Coccia A, Luparia P, Macrì L, Fiorito C, et al. Interpretation of p16INK4a/Ki-67 dual immunostaining for the triage of human papillomavirus-positive women by experts and nonexperts in cervical cytology. *Cancer Cytopathol* 2015; **123:** 212-8. doi: 10.1002/cncy.21511

4. McMenamin M, McKenna M, McDowell A, Dawson C, McKenna R. Intra- and inter-observer reproducibility of CINtecR PLUS in ThinPrep cytology preparations. *Cytopathology* 2017; **28:** 284-90. doi: 10.1111/cyt.12426

5. Benevolo M, Allia E, Gustinucci D, Rollo F, Bulletti S, Cesarini E, et al. Interobserver reproducibility of cytologic p16INK4a /Ki-67 dual immunostaining in human papillomavirus-positive women. *Cancer Cytopathol* 2017; **125:** 212-20. doi: 10.1002/cncy.21800

6. Kloboves Prevodnik V, Jerman T, Nolde N, Repše Fokter A, Jezeršek J, Pohar Marinšek Ž, et al. Interobserver variability and accuracy of p16/Ki-67 dual immunocytochemical staining on conventional cervical smears. *Diagn Pathol* 2019; **14:** 1-9. doi: 10.1186/s13000-019-0821-5

7. Ivanus U, Jerman T, Fokter AR, Takac I, Prevodnik VK, Marcec M, et al. Randomised trial of HPV self-sampling among non-attenders in the Slovenian cervical screening programme ZORA: comparing three different screening approaches. *Radiol Oncol* 2018; **52:** 399-412. doi: 10.2478/raon-2018-0036

8. R Core Team. *The R project for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019. [cited 2019 Dec 15]. Avalable at http://www.R-project.org/

9. National Cancer Institute. *Joinpoint Trend Analysis Software.* Joinpoint Regression Program, Version 4.6.0.0 - April 2018; Statistical Methodology and Applications Branch, Surveillance Research Program.

10. Altman DG. *Practical statistics for medical research*. Boca Raton: Chapman and Hall/CRC; 1991.