Original Article doi: 10.5146/tjpath.2017.01400

Diagnostic and Treatment Reproducibility of Cervical Intraepithelial Neoplasia / Squamous Intraepithelial Lesion and Factors Affecting the Diagnosis

Arzu SAĞLAM[1], Alp USUBÜTÜN[1], Anıl DOLGUN[2], George L. MUTTER[3], M. Coşkun SALMAN[4], Olcay KURTULAN[1], Aytekin AKYOL[1], Eylem AKAR ÖZKAN[5], Sema BAYKARA[6], Dilek BÜLBÜL[7], Zerrin CALAY[8], Funda EREN[9], Derya GÜMÜRDÜLÜ[10], Nihan HABERAL[5], Şennur İLVAN[8], Şeyda KARAVELİ[11], Meral KOYUNCUOĞLU[12], Bahar MÜEZZİNOĞLU[13], Kamil Hakan MÜFTÜOĞLU[14], Özlem ÖZEN[5], Necmettin ÖZDEMİR[15], Elif PEŞTERELİ[11], Çağnur ULUKUŞ[12], Osman ZEKİOĞLU[15]

Department of [1]Pathology and [2]Biostatistics, Hacettepe University, Medical Faculty, ANKARA, TURKEY, [3]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, [4]Department of Obstetrics & Gynecology, Hacettepe University, Medical Faculty, ANKARA, TURKEY Department of Pathology, [5]Baskent University, Medical Faculty, ANKARA, TURKEY, [6]Uludag University, Medical Faculty, BURSA, TURKEY, [7]Etlik Zubeyde Hanim Women's Health Teaching and Research Hospital, ANKARA, TURKEY, [8]Istanbul University Cerrahpasa, Medical Faculty, ISTANBUL, TURKEY, [9]Marmara University, Medical Faculty, ISTANBUL, TURKEY, [10]Cukurova University, Medical Faculty, ADANA, TURKEY, [11]Akdeniz University, Medical Faculty, ANTALYA, TURKEY, [12]Dokuz Eylül University, Medical Faculty, IZMIR, TURKEY, [13]Kocaeli University, Medical Faculty, KOCAELI, TURKEY, [14]Zekai Tahir Women's Health Teaching and Research Hospital, ANKARA, TURKEY, [15]Ege University, Medical Faculty, IZMIR, TURKEY

This study was presented as oral presentation at the 23th National Pathology Congress at İzmir, Turkey, held between 6-10 November 2013.

ABSTRACT

Objective: Inter-observer differences in the diagnosis of HPV related cervical lesions are problematic and response of gynecologists to these diagnostic entities is non-standardized. This study evaluated the diagnostic reproducibility of "cervical intraepithelial neoplasia" (CIN) and "squamous intraepithelial lesion" (SIL) diagnoses.

Material and Method: 19 pathologists evaluated 66 cases once using H&E slides and once with immunohistochemical studies (p16, Ki-67 and Pro-ExC). Management response to diagnoses was evaluated amongst 12 gynecologists. Pathologists and gynecologists were also given a questionnaire about how additional information like smear results and age modify diagnosis and management.

Results: We show moderate interobserver diagnostic reproducibility amongst pathologists. The overall kappa value was 0.50 and 0.59 using the CIN and SIL classifications respectively. Impact of immunohistochemical evaluation on interpretation of cases differed and there was lack of statistically significant improvement of interobserver diagnostic reproducibility with the addition of immunohistochemistry.

We saw that choice of treatment methods amongst gynecologists varied and overall concordance was only fair to moderate. The CIN2 diagnostic category was seen to have the lowest percentage agreement amongst both pathologists and gynecologists. We showed that pathologists had diagnostic "styles" and gynecologists had management "styles".

Conclusion: In summary each pathologist had different diagnostic tendencies which were affected not only by histopathology and marker studies, but also by the patient management tendencies of the gynecologist that the pathologist worked with. The two-tiered modified Bethesda system improved diagnostic agreement. We concluded that immunohistochemistry should be used only to resolve problems in select cases and not for every case.

Key Words: Interobserver reproducibility, SIL, CIN, Diagnosis, Gynecologist

INTRODUCTION

Diagnosis and management of Human Papilloma Virus (HPV)-related cervical lesions is a struggle. The main problem is which patient to treat, a decision largely (not solely) based on pathological diagnosis. Diagnosis is nontrivial due to conflicting classification schemas [3-class cervical intraepithelial neoplasia (CIN) vs. 2-class squamous intraepithelial lesion (SIL)] and subjective diagnostic criteria that are variously interpreted amongst pathologists (1-3). A recent study by Gage et al. showed that women would have a different probability of being treated depending on which laboratory and hence which pathologists reviewed the biopsy specimen (4).

The aim of this study was to assess the interobserver reproducibility of the two classification systems of HPV-related lesions of the cervix, namely the three-tiered "CIN"

(Turk Patoloji Derg 2017, 33:177-191)

Received : 10.04.2017 **Accepted :** 17.04.2017

Correspondence: Alp USUBÜTÜN
Hacettepe University, Medical Faculty, Department of Pathology, Sihhiye, Ankara, 06100 Turkey
E-mail: ausubutu@hacettepe.edu.tr Phone: +90 312 305 15 63

system and the two-tiered "Modified Bethesda" system (SIL) and to determine if there were any influences other than morphology on the diagnoses made.

Disease specific biomarkers, such as immunohistochemical (IHC) stains for p16, Ki-67 and Pro-ExC, have emerged as adjunctive tools for lesion classification. Shortfalls in assessing their utility include lack of a clear diagnostic gold standard, and uncertainty regarding when they should be implemented and how they are interpreted. We tackled some of these questions by measuring interobserver interpretive concordance for p16, Ki-67 and Pro-ExC, and benchmarking how they influenced diagnostic decision making.

Lastly, clinical management response to different diagnoses was evaluated amongst our gynecologist oncologists.

MATERIALS and METHODS

Case Selection

21 pathologists from 11 centers joined the study. Each center contributed six cervical biopsy cases for the study. The diagnostic spectrum included reactive, "low grade squamous intraepithelial lesion" (LSIL),"high grade squamous intraepithelial lesion" (HSIL) and microinvasive squamous cell carcinoma (mSCC). A total of 66 cases were collected (19 cervical biopsies, 44 LEEP/conization materials and 3 hysterectomy specimens). Only one representative slide from each case was selected.

Microscopic Examination

The pathologists assessed cases in two rounds, blinded to the original diagnosis and clinical features in each. They stratified all cases according to the CIN (CIN1, CIN2, CIN3) and SIL (LSIL, HSIL) classification systems with an additional group for the reactive and mSCCs. Round one was the "initial H&E round" where only H&E stained sections were evaluated. They also stated if they would require IHC studies to complement the diagnosis. The second round was the "follow-up with immunohistochemistry (IHC)" round, where cases were reevaluated along with IHC stains for p16, Ki-67 and Pro-ExC.

The IHC stains were scored in a three tiered pattern as detailed below (5):

P16: 1=negative, no or basal-only staining; 2=equivocal, bandlike staining of basal layer; 3=positive, full thickness staining

Ki67: 1=negative, <25% of cells stain; 2=equivocal, 25-50% of cells stain; 3=positive, >50% of cells stain.

ProExC: 1=negative, <25% of cells stain; 2=equivocal, 25-50% of cells stain; 3=positive, >50% of cells stain.

A total of 19 pathologists completed all phases of the study.

Questionnaire

Pathologists completed a questionnaire about factors that influence their diagnosis. Gynecologic oncologists from each contributing center were also queried with a questionnaire. They were asked to choose from 6 different treatment options present within these questionnaires as detailed below:

1- Therapy for infection

2- Follow-up with smear examinations

3- Follow-up with smear and colposcopy

4- Surface ablative therapy (laser or cryosurgery).

5- Conization

6- Hysterectomy

They completed the questionnaires twice, first, using the original (pre-study) pathology report, and second, a "re-edited standardized" post-study report, where all reports had the same format and all biopsy specimens accepted as LEEP material (so that differences in biopsy sizes like punch biopsy and hysterectomy would not be an additional confounding factor). Our goal was to assess factors that influenced gynecologic oncologists choice of treatment, and how patient management changed by pathologic diagnosis.

The questionnaires given to pathologists and gynecologic oncologist also contained questions regarding training and practice environment.

Statistical Analysis

Inter-observer reproducibility between the 19 reviewing pathologists was calculated using the kappa statistic (κ) for multiple raters when there are more than two diagnostic outcomes (6). The 95% bootstrap confidence intervals were calculated for the kappa statistics. The calculation was carried out separately for the two diagnostic rounds. The same calculation was repeated to assess the reproducibility of interpretation of the IHC stains. A consensus diagnosis was extracted for each case by using the majority-rule diagnoses of 19 different pathologists. Moreover, overall and category specific proportions of agreement (form raters) were calculated to assess the agreement of surveillance (options 1,2,3 above) compared to ablative or surgical (options 4,5,6 above) management preferences of the gynecologic oncologist. The kappa values were read as follows, 0: no agreement better than chance; 0-0.2: poor

agreement; 0.2-0.4: fair agreement; 0.4-0.6: moderate agreement; 0.6-0.8: substantial agreement; 0.8-1: almost perfect agreement (7). Mc-Nemar Bowker test was used to assess the differences in pathologist's classifications between the two rounds. Kappa analyses and the statistical tests were performed in STATA version 12.0 (StataCorp. Texas, USA). The statistical significance was set at $p<0.05$. Diagnostic trends were examined by hierarchical cluster analysis in a heat-map (color=diagnosis) matrix of reviewer by case (X and Y axis, respectively). For unsupervised hierarchical cluster analysis, euclidian distance measure was used, with Ward's linkage method performed in R (version 3.1.1, 2014) software. [R: A Language and Environment for Statistical Computing, author=R Core Team, R Foundation for Statistical Computing. Vienna, Austria, 2014. {https://www.R-project.org}]

The study has been approved by the institutional ethical committee (Hacettepe University Ethical committee, 5 June 2012, HEK 12/56-40).

RESULTS

Characteristics of the Pathologists

The 19 pathologists (Table I) were from university and community hospitals in different regions of Turkey with varying gynecologic workloads, duration of practice experience and practice context.

Factors That Influenced Pathologist Diagnoses According to the Questionnaire

Histopathology, IHC and smear results were most influential. The treatment preferences (ablation vs. surveillance) of the gynecologic oncologists the pathologists worked with, also had an effect on diagnoses rendered (Table II).

Interobserver Reproducibility of Diagnoses and Immunostain Interpretation

The inter-observer diagnostic concordance between the 19 pathologists for the "initial H&E" and "follow-up with IHC" rounds are summarized in Table III.

Table I: General characteristics of the participating pathologists and their agreement (weighted Kappa values**) with the majority-rule consensus diagnosis

	*Years of practice, experience	Practicing Hospital	Routine schema	CIN		SIL	
Pathologist				"initial HE round"	"follow-up with IHC"	"initial HE round"	"follow-up with IHC"
A	C / GynP	University	Both	0.89	0.77	0.86	0.81
B	B / GP	University	SIL	0.78	0.89	0.79	0.87
C	Resident	University	Both	0.66	0.63	0.58	0.63
D	C / GynP	University	Both	0.86	0.73	0.85	0.73
E	B / GP	University	SIL	0.77	0.76	0.75	0.84
F	C / GynP	University	Both	0.97	0.88	0.99	0.89
G	C / GynP	University	SIL	0.79	0.71	0.80	0.71
H	C / GynP	University	Both	0.92	0.89	0.92	0.89
I	C / GynP	University	Both	0.72	0.74	0.81	0.78
J	C / GynP	University	Both	0.86	0.78	0.85	0.74
K	C / GynP	Community	SIL	0.65	0.69	0.72	0.78
L	C / GynP	Community	Both	0.83	0.83	0.87	0.84
M	B / GynP	University	Both	0.82	0.81	0.82	0.79
N	B/ GP	University	Both	0.72	0.74	0.69	0.71
O	C /GynP	University	Both	0.74	0.79	0.78	0.76
P	C / GynP	University	Both	0.74	0.84	0.74	0.77
Q	C / GynP	University	Both	0.73	0.75	0.78	0.83
R	C / GynP	University	Both	0.71	0.88	0.72	0.87
S	C / GynP	University	CIN	0.79	0.85	0.78	0.86

$p<0.001$, **GP:** General pathologist, **GynP:** Pathologist with experience in gynecologic pathology, **IHC:** Immunohistochemistry. *(Years of practice A=0-3 years, B=3-10 years, C=more than 10 years) Kappa values ranged from 0.69 to 0.99), with the exception of one outlier, a resident in training, was noted to have the lowest kappa values of 0.58-0.66.

The agreement was moderate with both classification systems, the SIL classification system having a higher kappa value. IHC evaluation did not significantly improve inter-observer diagnostic reproducibility within either classification system ($p<0.05$ both for CIN and SIL).

A majority-rules consensus was calculated for each case during each round. Inter-observer reproducibility (weighted Kappa values) of the pathologists, with regard to the majority-rule consensus diagnosis ranged from 0.69 to 0.99, with the exception of one outlier, a resident in training, who had the lowest kappa values of 0.58-0.66 (Table I).

SIL and CIN consensus diagnoses of the cases for the first and second round were cross-matched (Table IV, V) except for one case all CIN2-3 were HSIL and all CIN1 were LSIL.

Overall kappa values (interobserver reproducibility) amongst the 19 pathologists for interpretation of each individual IHC stain and the kappa values with regard to each score are given in Table VI. There was a moderate to substantial agreement in interpretation of IHC with judgment of score 2 being the most problematic.

Individual pathologists displayed different diagnostic patterns. For example, some stood out by high percentage of use of certain categories such as CIN2. This can be seen in Figures 1 and 2. Two major diagnostic styles emerged in which membership was highly conserved (17/19) by diagnostic schema used. Generally, the rightmost diagnostic style group had a tendency to push SIL and CIN diagnoses to a higher grade – a diagnostically aggressive group (tendency to upgrade –"upG"), whereas the left most group tended to do the opposite (tendency to down grade – "downG").

Table II: Factors that affect diagnostic decision making for the pathologist

Affecting Factors	Never (%)	Rarely (%)	Sometimes (%)	Often (%)	Always (%)
Patient age	9.5	23.8	47.6	9.5	9.5
Clinical diagnostic impression	19	19.0	38.1	19.0	4.8
Gynecologic oncologist treatment preferences	9.5	19.0	33.3	28.6	9.5
Histopathology	-	-	4.8	9.5	87.5
Pap smear results, concurrent and/or prior to biopsy	-	9.5	28.6	42.9	19.0
p16	14.3	4.8	23.8	23.8	33.3
Ki-67	14.3	4.8	28.6	38.1	14.3
ProExC	90.0	-	5.0	-	5.0
HPV DNA status	25.0	10.0	20.0	35.0	10.0

Table III: Inter-observer diagnostic reproducibility between the 19 pathologists for the "initial HE" and "follow-up with immunos" rounds for the CIN and SIL classification systems

		"initial HE round" KAPPA	"follow-up with IHC"* KAPPA
CIN	Reactive	0.66	0.67
	CIN1	0.49	0.53
	CIN2	0.24	0.22
	CIN3	0.49	0.51
	mSCC	0.55	0.54
	Overall	0.50[§]	0.50[¥]
SIL	Reactive	0.65	0.67
	LSIL	0.49	0.53
	HSIL	0.62	0.62
	mSCC	0.54	0.54
	Overall	0.59[£]	0.60[ǂ]

All Kappa values were statistically significant ($p<0.001$).

*Together with immunohistochemistry.

[§] 95% bootstrap confidence interval for the overall Kappa: (0.437 - 0.552), [¥] 95% bootstrap confidence interval for the overall Kappa: (0.460 - 0.549).

[£] 95% bootstrap confidence interval for the overall Kappa: (0.548 - 0.631), [ǂ] 95% bootstrap confidence interval for the overall Kappa: (0.557 - 0.648).

Diagnostic styles of individual pathologists was mostly conserved across diagnostic schema (CIN to SIL) (Table VII). "Initial H&E" round to "follow-up with IHC" round crossover of individual pathologists from one diagnostic style group to another however occurred with equal frequency in both directions: 50% (3/6) downG to upG, 50% (5/10) upG to downG. It seems likely that individuals were affected in a different manner by IHC.

Diagnostic Impact of Immunohistochemistry

Diagnostic changes made by pathologists after IHC and its impact on inter-observer reproducibility were not statistically significant (Table I, $p<0.05$), but we can identify several trends. IHC improved segregation of cases into specific diagnostic groups when compared to H&E review alone. This is evident as increased homogeneity of the horizontal rows (cases) of the heat maps in Figures 1 and 2. A decline in use of CIN2 diagnoses in the "follow-up with IHC" round, with increased frequency of diagnosis of CIN3 and HSIL polarized the categories more strongly. Interestingly the diagnosis of mSCC decreased after IHC evaluation, as areas suspicious of microinvasion on H&E turned out to be glandular involvement made clear by serial sectioning and highlighting of the epithelial-stromal interface.

Five pathologists made significant changes in their diagnoses after the addition of IHC, including two not experienced in gynecologic pathology, and three gynecologic pathologists.

Unblinded Re-Review of Most Discordant Cases

Five cases in which more than half the pathologists stated that they would order IHC turned out to be the ones in which most diagnostic change was made between the two rounds. Examination of these cases (Table VIII) revealed that some had areas where the differential diagnosis of benign lesions like inflammation associated changes had to be entertained (cases 6 and 10). Case 6 is characteristic; before IHC except for one, all "downG" group pathologists diagnosed it as reactive while "upG" group pathologists as HSIL/mSCC. After IHC the diagnosis was HSIL or mSCC by both groups of pathologist (Figure 3A-D).

Table IV: Comparison of CIN and SIL consensus diagnoses in the "initial HE round"

	SIL "initial HE round" Consensus Diagnoses				
CIN "initial HE round" Consensus Diagnoses	**Reactive**	**LSIL**	**HSIL**	**mSCC**	**Total**
Reactive	16	0	0	0	16
CIN1	0	16	0	0	16
CIN2	0	0	7	0	7
CIN3	0	0	20	0	20
mSCC	0	0	1	6	7
Total	16	16	28	6	66

Table V. Comparison of CIN and SIL consensus diagnoses in the "follow-up with IHC" round

	SIL "follow-up with immunos" Consensus Diagnoses				
CIN "follow-up with immunos" Consensus Diagnoses	**Reactive**	**LSIL**	**HSIL**	**mSCC**	**Total**
Reactive	15	0	0	0	15
CIN1	0	12	1	0	13
CIN2	0	0	6	0	6
CIN3	0	0	26	0	26
mSCC	0	0	0	5	5
Total	15	12	33	5	65

Table VI: Kappa values of interpretation of immunohistochemical staining.

Stain Score	Ki-67	ProExC	p16
1	0.70	0.70	0.76
2	0.23	0.35	0.38
3	0.63	0.74	0.65
Overall	0.54	0.63	0.62

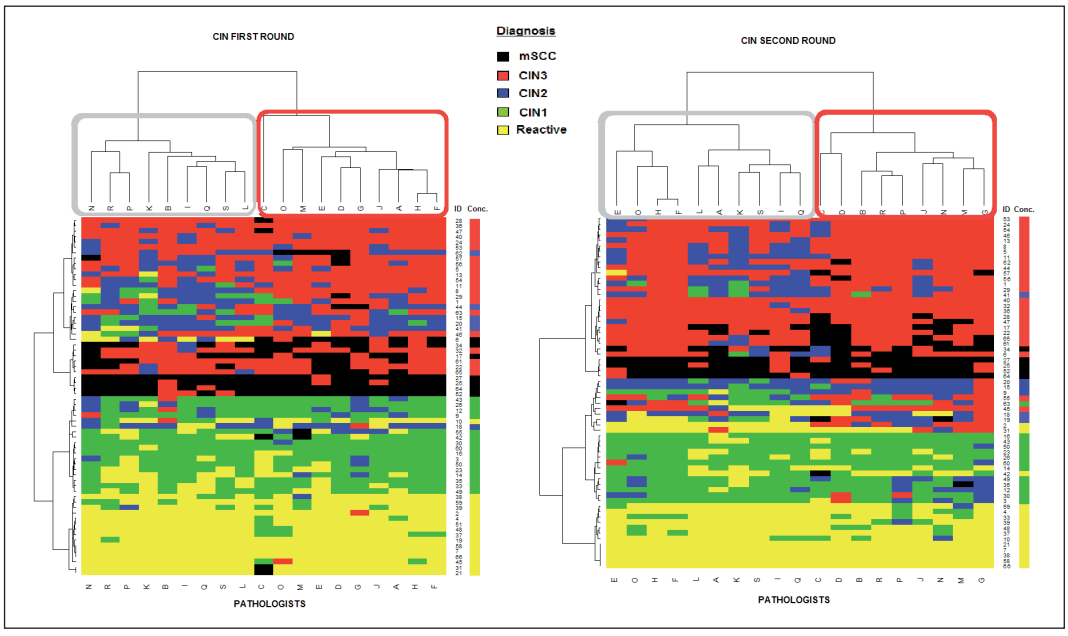

Figure 1: Heat map demonstrating unsupervised clustering of CIN diagnoses (color) by the reviewing pathologist (columns) and individual cases (rows). Left panel diagnoses are based on the "initial H&E" round, and right panel diagnoses are based on the "follow-up with IHC" round (diagnoses rendered using p16, Pro-ExC and Ki67). Addition of IHC in the second round improved consistency of distinction across two major diagnostic thresholds: 1) reactive (yellow) vs. CIN1 (green) lesions; and 2) reactive (yellow) vs. CIN3 (red) lesions. This is seen as greater consistency between pathologists for these diagnoses (rows more homogenous) in the right panel. Pathologist diagnostic style groups according to diagnoses is shown by major node separation in the tree above the heat maps (pathologist clusters, major nodes to left= "gray" and right= "red"). The detached heat column to the side of each figure shows the majority-rule consensus diagnosis for each case.

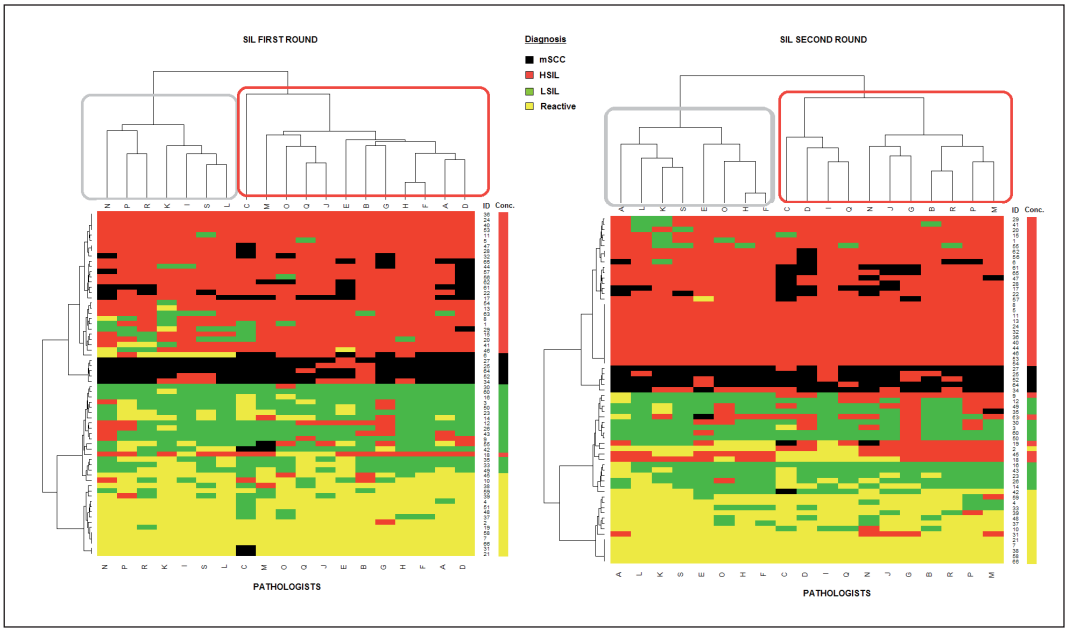

Figure 2: Heat map demonstrating unsupervised clustering of SIL diagnoses (color) by reviewing pathologist (columns) and individual specimens (rows). Left panel diagnoses are based only on the "initial H&E"round, and right panel diagnoses are rendered using H&E plus p16, Pro-ExC and Ki67 IHC stains. The detached heat column to the side of each figure shows the majority-rule consensus diagnosis for each case. As with the CIN classification (Figure 1), addition of IHC in the "follow-up with immunos" round improved consistency of distinction across major diagnostic thresholds. Pathologist diagnostic style groups according to diagnoses is shown by major node separation in the tree above the heat maps (pathologist clusters, major nodes to left= "gray" and right= "red").

In others the problem was differentiation of koilocytosis versus superficial vacuolization (cases 33 and 39) and differentiation of LSIL from HSIL was the challenge (case 44, Figure 4A-D). Within this group, addition of IHC (combined interpretation of all 3 markers) reduced diagnostic discordance. Positive IHC tended to increase, whereas negative IHC tended to decrease the grade of the lesion.

Cases 2, 19 and 45, which were accompanied by severe inflammation were diagnosed as reactive (consensus diagnosis) in the "initial H&E" round by both ("downG" and "upG") groups. After positive IHC, the consensus diagnosis for cases 19 and 45 was HSIL/mSCC and for case 2 the consensus diagnosis was reactive although almost half diagnosed it as HSIL (Figure 5A-D). Furthermore some cases were diagnosed as LSIL in the "initial H&E" round but HSIL after IHC by pathologists in the "upG" group; however during re-review we thought that some of these cases actually lacked decisive IHC staining that would lead to their upgrading (Figure 6A-D). Such cases emphasized the impact of "diagnostic styles" on overall IHC interpretation.

Characteristics of the Participating Gynecologic Oncologists

The 12 gynecologic oncologists were from university and community hospitals in different regions of Turkey. They had varying workloads and differed in the duration of practice experience and practice context. They reported histology, smear results and patient's age to be most influential on diagnostic decision making (Table IX).

Interobserver Reproducibility of Patient Management Among Gynecologic Oncologists

Concordance of treatment methods amongst gynecologic oncologists for the patient group was only fair (kappa value: 0.2974, data not shown). When the management categories were reduced to three as noninvasive (infection therapy + follow-up with smear examinations + follow-up with smear and colposcopy), ablative (destruction and conization) and hysterectomy, the overall kappa value reached moderate levels (0.57) (Table Xa). The CIN2 diagnostic category was seen to have the lowest percentage agreement, whereas reactive and CIN1/SIL had the highest agreement (Table Xb,c).

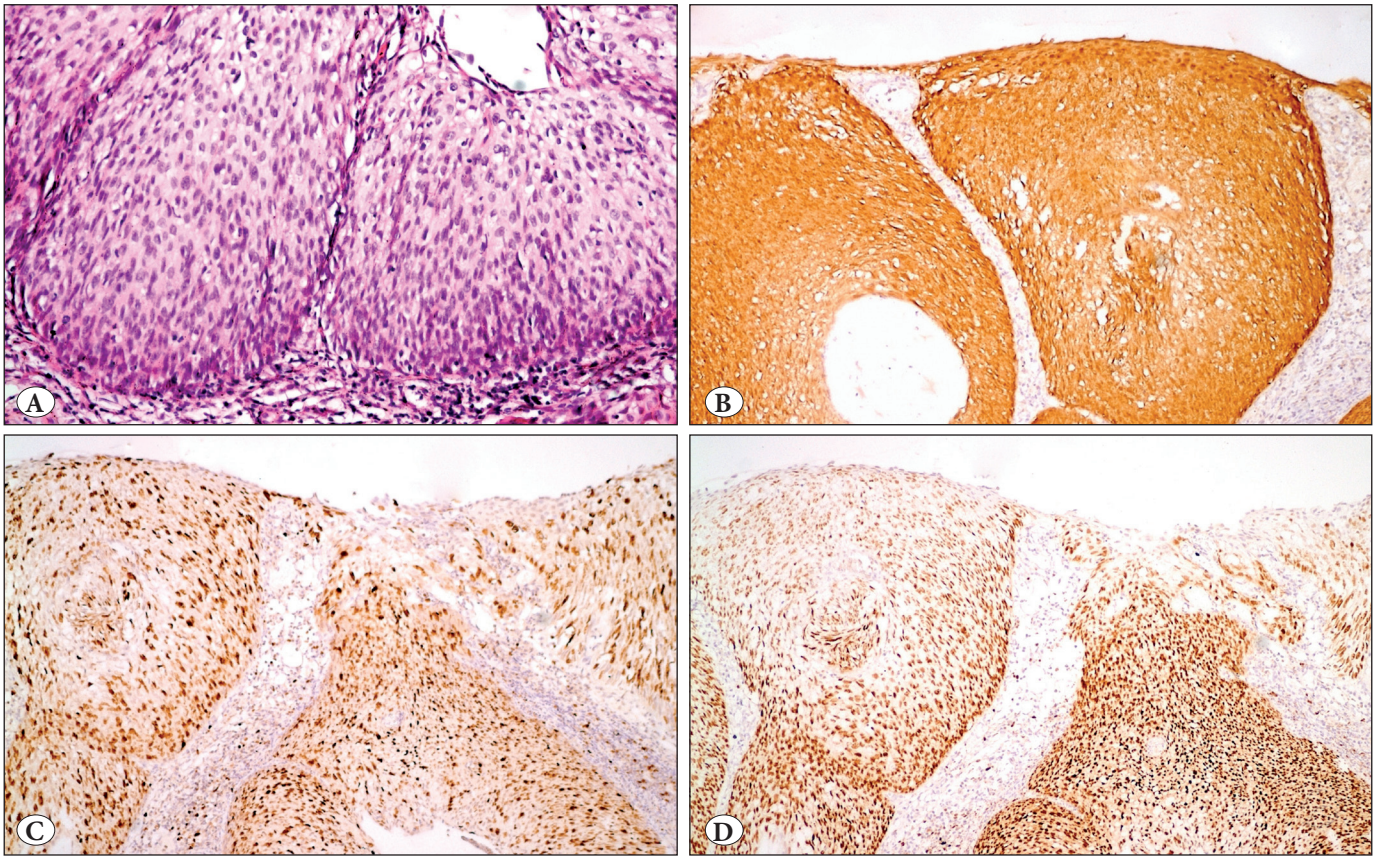

Figure 3: Case 6, a case with a diagnostic challenge of reactive changes (favored by the "downG" group) vs. CIN3/HSIL (favored by the "upG" group) during the initial H&E round. The discrepancy was resolved after the "follow-up with immunohistochemistry round" where the diagnosis was HSIL or mSCC by both groups of pathologist (**A:** H&E; x400, **B:** p16; x400, **C:** Ki-67; x200 , **D:** Pro-ExC; x200).

Table VII: Changes between reads ["initial HE"(R1) vs. "follow-up with IHC"(R2)] in diagnostic style group "downG" (Gray cluster in heat map) or "upG" (Red cluster in heat map) of pathologists based on hierarchical clustering of pathologists in Figures 1 and 2

Pathologist	CINR1	SILR1	CINR2	SILR2	"downG" to "upG" Group Switch
N	downG	downG	upG	upG	downG to upG
P	downG	downG	upG	upG	downG to upG
R	downG	downG	upG	upG	downG to upG
I	downG	downG	downG	upG	Partial
K	downG	downG	downG	downG	no
L	downG	downG	downG	downG	no
S	downG	downG	downG	downG	no
B	downG	upG	upG	upG	Partial
Q	downG	upG	downG	upG	Partial
C	upG	upG	upG	upG	no
D	upG	upG	upG	upG	no
G	upG	upG	upG	upG	no
J	upG	upG	upG	upG	no
M	upG	upG	upG	upG	no
A	upG	upG	downG	downG	upG to downG
E	upG	upG	downG	downG	upG to downG
F	upG	upG	downG	downG	upG to downG
H	upG	upG	downG	downG	upG to downG
O	upG	upG	downG	downG	upG to downG

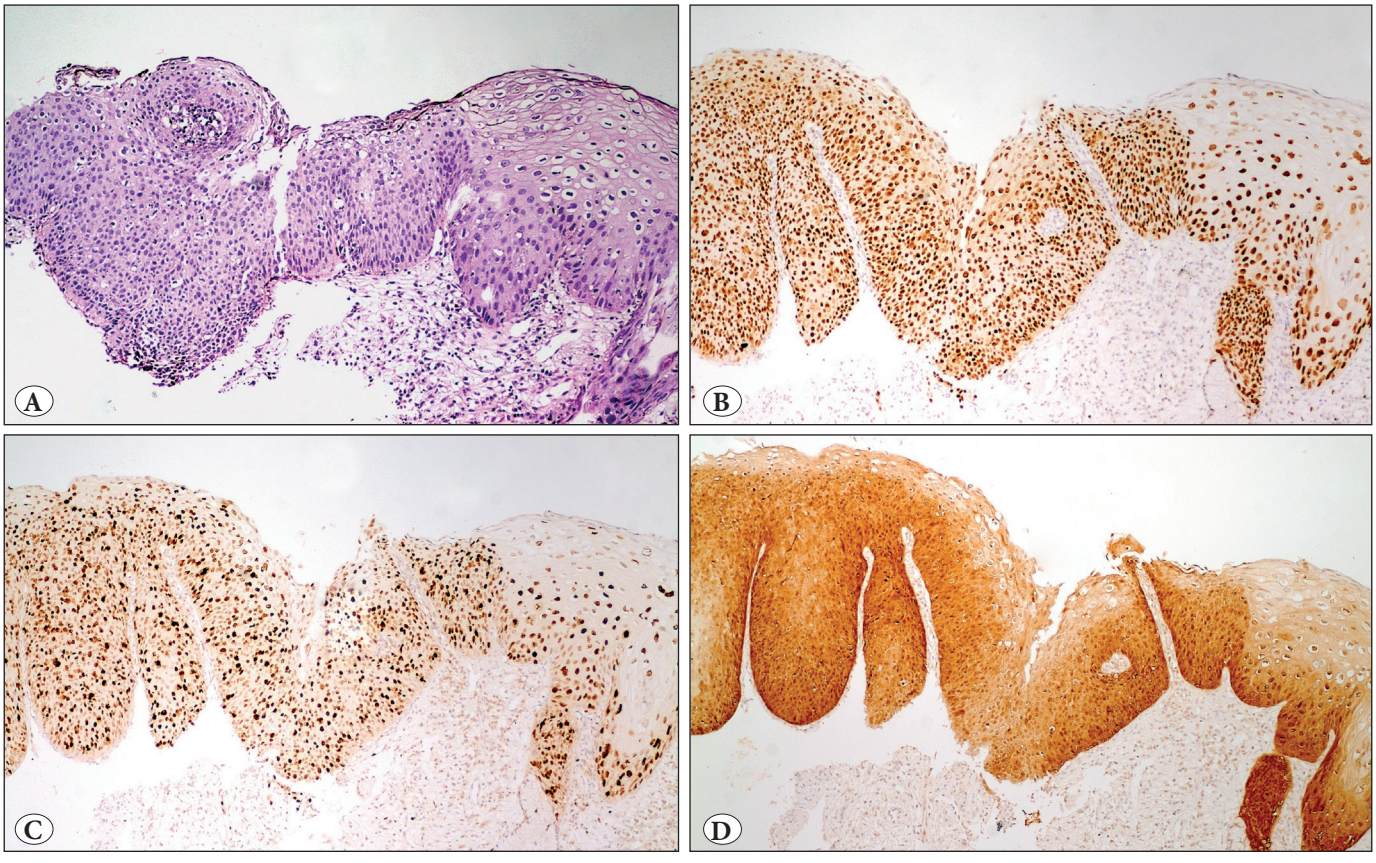

Figure 4: Case 44, demonstrating challenge in differentiation of LSIL from HSIL.
(**A:** H&E; x100, **B:** Pro-ExC; x100, **C:** Ki-67; x100, **D:** p16; x100).

Table VIII: Diagnostic spectrum of 19 reporting pathologists for the most discordant cases

	CIN				Mod. Bethesda (SIL)				IHC	Comment*
	Before IHC %		After IHC %		Before IHC %		After IHC %			
Case 6	Reactive CIN2 CIN3 mSCC	31.6 15.8 15.8 36.8	CIN1 CIN2 CIN3 mSCC	5.3 15.8 57.9 21.1	Reactive HSIL mSCC	31.6 31.6 36.8	LSIL HSIL mSCC	5.3 73.7 21.1	p16 pos Ki-67 pos Pro-ExC pos	HSIL
Case 10	Reactive CIN1 CIN2 CIN3	57.9 26.3 10.5 5.3	Reactive CIN1 CIN2	73.7 21.1 5.3	Reactive LSIL HSIL	57.9 26.3 15.8	Reactive LSIL HSIL	73.7 21.1 5.3	p16 neg Ki-67 neg Pro-ExC neg	Reactive
Case 33	Reactive CIN1	31.6 68.4	Reactive CIN1	68.4 31.6	Reactive LSIL	31.6 68.4	Reactive LSIL	68.4 31.6	p16 neg Ki-67 neg Pro-ExC neg	Reactive
Case 39	Reactive CIN1 CIN2	68.4 26.3 5.3	Reactive CIN1 CIN2	84.2 10.5 5.3	Reactive LSIL HSIL	68.4 26.3 5.3	Reactive LSIL HSIL	84.2 10.5 5.3	p16 neg Ki-67 neg Pro-ExC neg	Reactive
Case 44	CIN1 CIN2 CIN3 mSCC	10.5 52.6 26.3 10.5	CIN2 CIN3	26.3 73.7	Reactive LSIL HSIL	10.5 10.5 78.9	HSIL	100.0	p16 pos Ki-67 pos Pro-ExC pos	HSIL

* Unblinded consensus comments by the pathologist who designed the study and his group. (**IHC:** Immunohistochemistry).

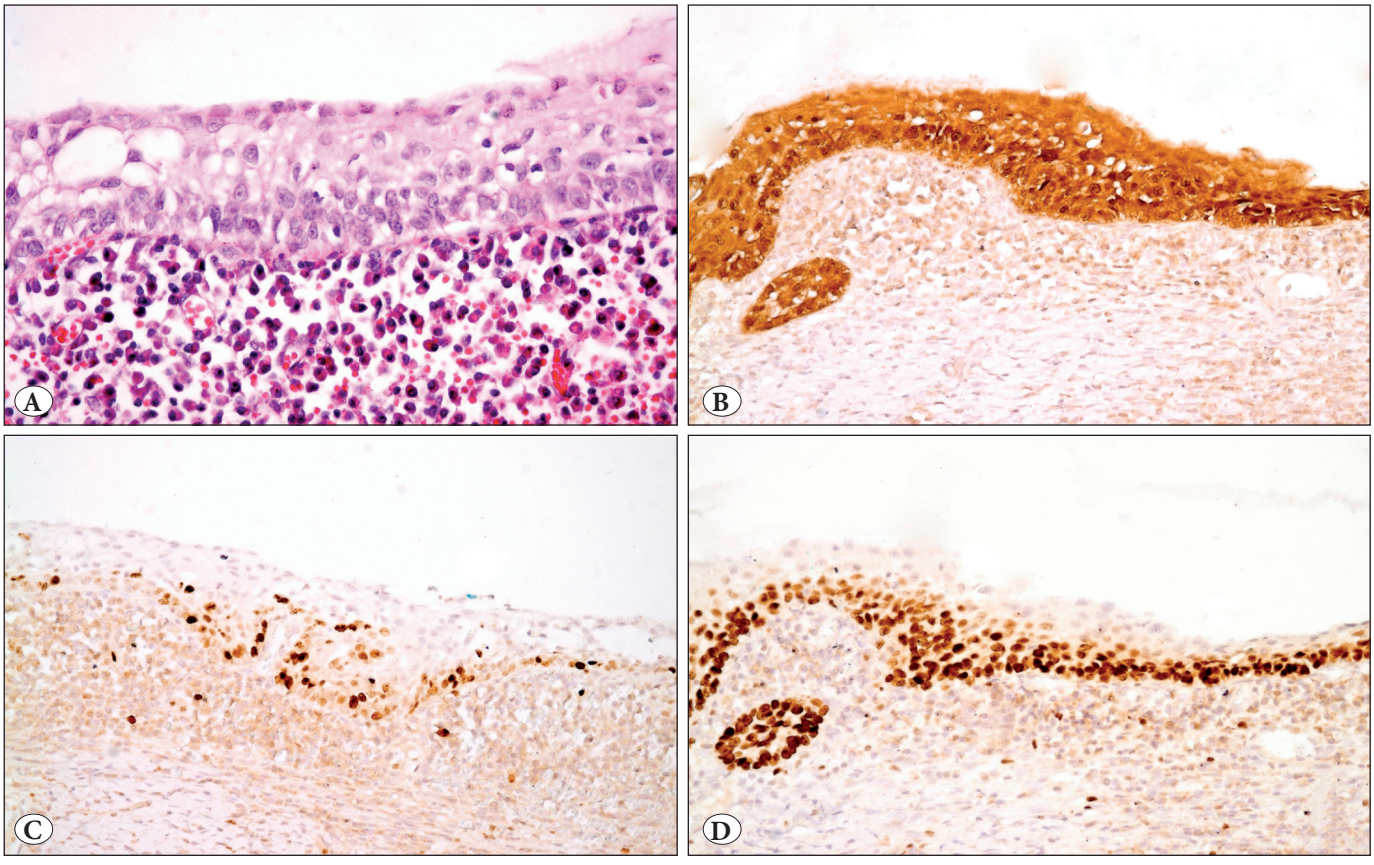

Figure 5: Case 2, a case with accompanying inflammation, diagnosed as HSIL by almost half the pathologists, and the consensus diagnosis was reactive (**A:** H&E; x200, **B:** p16; x200, **C:** Ki-67; x200, **D:** Pro-ExC; x200).

Kappa values did not differ significantly after the gynecologic oncologists were given re-edited standardized reports for all patients during the second round (Table Xa).

As with the pathologists, individual gynecologic oncologists displayed different management styles and clustered in two groups (Figure 7). Generally, the rightmost (RED) management style had a higher tendency of ablative and surgical treatment – a therapeutically aggressive group.

Table XI summarizes the consensus management decisions with regard to diagnostic categories. All reactive and CIN1/LSIL cases were assigned to the non-invasive therapy group whereas therapy options varied more widely with CIN2/CIN3/HSIL diagnoses. When management decisions are analyzed on a case-by-case basis it can easily be recognized that the management of some cases was incompatible with the general tendency (Figure 7 and Table XII). In cases for whom the consensus management was noninvasive, hysterectomy might be preferred due to coinciding conditions necessitating hysterectomy. For the two patients with a diagnosis of microinvasive carcinoma choice of a noninvasive management may be explained by the fact that both patients were young (desire for children?). Since information pertaining to marital status, parity and fertility desire was not obtained during the study and hence provided to the gynecologic oncologists, one can only speculate.

Table IX -Factors that affect gynecologic oncologist management

Factor	Never (%)	Rarely (%)	Sometimes (%)	Often (%)	Always (%)
Patients age	-	-	-	16.7	83.3
Histopathology	-	-	-	-	100
PAP Cytology	-	-	8.3	41.7	50
HPV DNA status	8.3	33.3	16.7	33.3	8.3

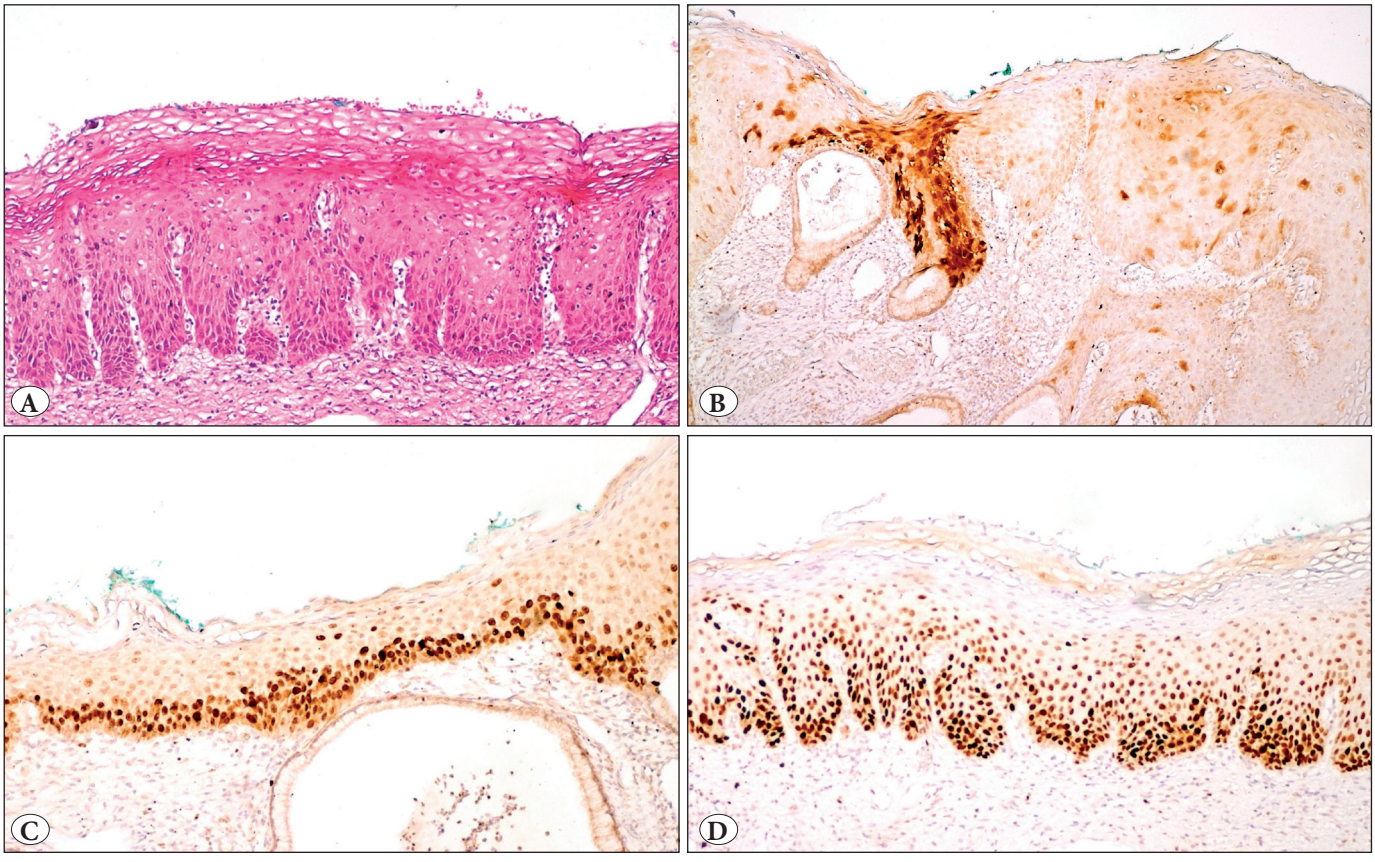

Figure 6: Case 35, diagnosed as LSIL in the "initial H&E "round but HSIL after immunohistochemical study. (**A:** H&E; x100, **B:** p16; x100, **C:** Ki-67; x100, **D:** Pro-ExC; x100); immunohistochemistry is ambiguous.

Table Xa: Interobserver reproducibility of patient management between gynecologic oncologists

Therapy Options	KAPPA value "1st round"	KAPPA value "2nd round"
Non-invasive	0.68	0.68
Ablative	0.48	0.44
Hysterectomy	0.48	0.52
Overall	0.57[¥]	0.57[$]

All Kappa values were statistically significant (p<0.05)

[¥] 95% Confidence Interval for the first round overall Kappa (0.533 - 0.626), [$] 95% Confidence Interval for the second round overall Kappa (0.528 - 0.579)

Table Xb: Agreement on therapeutic management with regard to CIN diagnostic categories

Diagnostic Categories	n	P(1)	P(2)	P(3)	Po
Reactive	9	0.95	0.18	0.00	0.91
CIN1	20	0.96	0.00	0.09	0.93
CIN2	4	0.60	0.59	0.27	0.57
CIN3	26	0.74	0.68	0.33	0.66
mSCC	7	0.12	0.30	0.75	0.62
Overall	66	0.88	0.61	0.55	0.77

P(1): Percentage of agreement for the non-invasive treatment category, **P(2):** Percentage of agreement for the ablative treatment category.
P(3): Percentage of agreement for the hysterectomy treatment category, **Po:** Overall percentage of agreement for all therapeutic categories.

Table Xc: Agreement on therapeutic management with regard to SIL diagnostic categories

Diagnostic Categories	n	P(1)	P(2)	P(3)	Po
Reactive	9	0.95	0.18	0.00	0.91
LSIL	20	0.96	0.00	0.09	0.93
HSIL	30	0.72	0.67	0.30	0.60
mSCC	7	0.12	0.30	0.75	0.62
Overall	66	0.88	0.60	0.55	0.77

P(1): Percentage of agreement for the non-invasive treatment category, **P(2):** Percentage of agreement for the ablative treatment category.
P(3): Percentage of agreement for the hysterectomy treatment category, **Po:** Overall percentage of agreement for all therapeutic categories.

Table XI: Majority-rule consensus management option with regard to the CIN and SIL diagnostic categories

Original Diagnosis	Majority-rule consensus of therapy options			
	Non-invasive	Ablative	Hysterectomy	Total
Reactive	9 (100%)	-	-	9
CIN1	20 (100%)	-	-	20
CIN2	2 (50%)	2 (50%)	-	4
CIN3	9 (34.6%)	16 (61.5%)	1 (3.8%)	26
LSIL	20 (100%)	-	-	20
HSIL	11 (36.7%)	18 (60%)	1 (3.3%)	30
mSCC	-	1 (14.3%)	6 (85.7%)	7
Total	40 (60.6%)	19 (28.8%)	7 (10.6%)	66

HSIL: High-grade Squamous Intraepithelial Lesion, **LSIL:** Low-grade Squamous Intraepithelial Lesion, **CIN:** Cervical Intraepithelial Neoplasia, **SCC:** Squamous Cell Carcinoma.

Table XII: Characteristics of cases whose managements were substantially incompatible with the consensus approaches.

Case no.	Age	Procedure	Diagnosis	Consensus approach	Incompatible approach
21	57	Biopsy	Chronic cervicitis	Noninvasive	Hysterectomy
28	65	LEEP	HSIL (CIN3) with intact borders	Noninvasive	Hysterectomy
33	62	LEEP	LSIL (CIN1) with intact borders	Noninvasive	Hysterectomy
46	37	LEEP	HSIL (CIN3) with intact borders	Noninvasive	Hysterectomy
58	67	LEEP	LSIL (CIN1) with involved borders	Noninvasive	Hysterectomy
52	35	LEEP	Microinvasive SCC	Hysterectomy	Noninvasive
65	35	LEEP	Microinvasive SCC	Hysterectomy	Noninvasive

LEEP: Loop Electrosurgical Excision Procedure, **HSIL:** High-grade Squamous Intraepithelial Lesion, **LSIL:** Low-grade Squamous Intraepithelial Lesion, **CIN:** Cervical Intraepithelial Neoplasia, **SCC:** Squamous Cell Carcinoma.

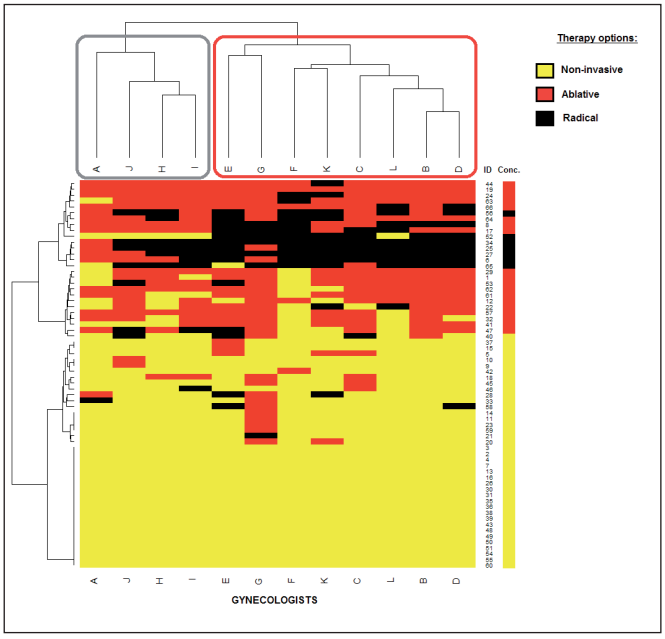

Figure 7: Heat map demonstrating unsupervised clustering of therapy options (color) by gynecologist (columns) and individual specimens (rows). Management style groups (grey=left, red=right) according to diagnoses is shown by major node separation in the tree above the heat maps. The detached heat column to the side of the figure shows the "Majority-rule consensus therapy option" for each case.

DISCUSSION

We evaluated diagnostic reproducibility of cervical SIL and CIN diagnoses, and explored factors that may modify diagnosis and therapeutic decisions. We measured the impact of IHC on diagnosis, and queried pathologists and gynecologic oncologist about how additional information such as smear results, age, modify diagnosis and management.

We show moderate interobserver diagnostic reproducibility by a mixed group of 19 pathologists evaluating HPV related lesions of the cervix. Our results are in accordance with previously reported CIN diagnosis inter-observer reproducibility's ranging from poor to good (kappa 0.23-0.64) (2, 3, 7-20) and likewise CIN 2 has the lowest interobserver diagnostic reproducibility (1-4, 7-11, 13-15, 21-23). Some pathologists who report results in the CIN system have reduced used of the CIN2 diagnostic category to such a low frequency that in their hands it becomes a *de facto* 2-class system. We saw this effect amongst some of our pathologists, where the frequency of use of CIN2 diagnoses ranged between 3 to 20% (one fifth) of cases. With the "Modified Bethesda" system the reduction of number of categories slightly improved reproducibility.

There are many study design factors that can influence measurements of diagnostic reproducibility. The spectrum of lesions included, sampling format, diagnostic schema employed, and number of reviewing pathologists are all contributors to kappa values reported (7, 9, 11, 17, 20). Subspecialty expertise does not necessarily enhance diagnostic consensus (23), a conclusion partially confirmed by us. Not having completed pathology training however was seen to impact diagnostic decision, since the pathology resident amongst our pathologists displayed the lowest agreement with respect to the consensus diagnosis. Inclusion of a large cohort of reviewing pathologists in our study can be expected to modulate the impact of outlier diagnostic behavior, and thus better approximate overall community patterns.

We noted that pathologists generally used the same criteria for assessing the cases whether they were to classify them as CIN or SIL, and hence the use of CIN versus SIL on a case

by case basis was generally compatible, almost all CIN1's were LSIL and CIN 2 and 3's were HSIL.

The potential benefit of IHC as an aid to improving diagnostic reproducibility was measured by comparison of diagnostic performance with and without the IHC stains. There was lack of statistically significant improvement of interobserver diagnostic reproducibility with the addition of IHC, contradictory to findings in the literature (11, 17, 18, 24-26). The confounding effect of IHC was less pronounced with the use of the Modified Bethesda classification.

According to the literature addition of p16 improves interobserver agreement (20), by pinpointing small lesions or highlighting lesions complicated by inflammation, as perfectly exemplified in two of our case which were diagnosed as reactive in this study by almost all participants in the "initial HE" round but changed to HSIL diagnosis after IHC.

A problem with all of these markers is that they are more useful in distinction between HPV related and non-viral (reactive or atrophy) lesions, but are less effective in differentiating between viral subsets of low grade and high grade lesions (27). In our study, the use of IHC was only helpful in a small number of cases and our results showed that the diagnosis tends to be upgraded with the use of IHC. We hence conclude that IHC should not be ordered for every case, but confined to those cases which are diagnostically ambiguous on H&E. We and others (27, 28), have stressed the risk of overtreatment which occurs when upgrading lesions with routine use of p16.

When the five least reproducible cases in our study were further evaluated these cases were seen to have elicited the highest rates of request from reviewing pathologists for IHC studies. Addition of IHC clearly helped resolve these problematic cases. It is important to note that combined interpretation of all three markers was able to achieve this result and detailed review of these cases showed that no marker by itself would have been sufficient.

Moreover we saw that choice of treatment methods amongst our 12 gynecologic oncologists for the same cases also varied and overall concordance was only fair and the kappa value merely increased to moderate with minimization of management categories. There was high agreement between gynecologic oncologists regarding management of reactive/low grade lesions, good agreement with respect to high grade lesions (HSIL, CIN3 and mSCC) and moderate agreement with CIN2. As with the pathologists the CIN2 diagnostic category had the lowest percentage agreement. The format/style of the pathology report did not influence the gynecologic oncologist's decision. Recommended management options for these lesions are clearly defined by guidelines which are widely recognized and accepted by Turkish gynecologists (29). Treatment variance may be a reflection of the role of institutional practice patterns and personal experience of the gynecologist. It could however also be a reflection of other confounding factors, such as patient compliance, fertility desire, age and patient preferences. Unfortunately, we were unable to assess these factors as covariates, as this information was not available. To our knowledge there is no other study in the English literature that analyzes the interobserver reproducibility of gynecologic oncologists with regard to management of patients with the same diagnosis and is a unique contribution of our study that deserves further expanded and in depth analysis.

We also queried pathologists for factors that influenced histologic diagnosis, and found cytology results and IHC were incorporated in the diagnostic process. One surprising diagnostic modifier was the differing management styles of the gynecologists the pathologists worked with. Presumably the pathologists were modifying diagnostic thresholds to accommodate differing risks of these reflex treatments by particular gynecologist oncologists.

We showed that pathologists had diagnostic "styles" (30). This is shown in Figure 1 where pathologists fell into two style groups: one had a tendency to push SIL and CIN diagnoses to a higher grade – a diagnostically aggressive group, whereas the other was more conservative. These styles were generally preserved irrespective of the classification system used (CIN or SIL), which shows that diagnostic behavior of the individual pathologist is not subject to change by simple replacement of terminology. Interestingly though some of the pathologists' diagnostic styles changed following IHC. It therefore seems likely that IHC findings may modify the diagnostic style of pathologists. On the other hand, diagnostic style may modify IHC interpretation and its impact on diagnosis.

In summary, both the diagnosis and clinical management of cervical HPV lesions is problematic. Appropriate patient management is not merely pure morphologic assessment and may be influenced by factors that are hard to clarify. As more data on clinical follow-up of problematic cases accumulate and stricter and objective criteria that help classify cases into those that will or will not progress come out, these problems may be better resolved.

CONFLICT of INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENT

We thank Sitogen and BD for providing the IHC markers for the study and we thank Özlem Kalaycı for her technical assistance.

REFERENCES

1. McCluggage WG, Walsh MY, Thornton CM, Hamilton PW, Date A, Caughley LM, Bharucha H. Inter- and intra-observer variation in the histopathological reporting of cervical squamous intraepithelial lesions using a modified Bethesda grading system. Br J Obstet Gynaecol. 1998;105:206-10.
2. Basu P, Kamal M, Ray C, Bhat D, Ghosh I, Mittal S, Chatterjee S, Samaddar A, Biswas J. Interobserver agreement in the reporting of cervical biopsy specimens obtained from women screened by visual inspection with acetic acid and hybrid capture 2. Int J Gynecol Pathol. 2013;32:509-15.
3. McCluggage WG, Bharucha H, Caughley LM, Date A, Hamilton PW, Thornton CM, Walsh MY. Interobserver variation in the reporting of cervical colposcopic biopsy specimens: Comparison of grading systems. J Clin Pathol. 1996;49:833-5.
4. Gage JC, Schiffman M, Hunt WC, Joste N, Ghosh A, Wentzensen N, Wheeler CM. Cervical histopathology variability among laboratories: A population-based statewide investigation. Am J Clin Pathol. 2013;139:330-5.
5. Walts AE, Bose S. p16, Ki-67, and BD ProExC immunostaining: A practical approach for diagnosis of cervical intraepithelial neoplasia. Hum Pathol. 2009;40:957-64.
6. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. New York: John Wiley&Sons, 2003.
7. 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.
8. de Vet HC, Knipschild PG, Schouten HJ, Koudstaal J, Kwee WS, Willebrand D, Sturmans F, Arends JW. Interobserver variation in histopathological grading of cervical dysplasia. J Clin Epidemiol. 1990;43:1395-8.
9. de Vet HC, Knipschild PG, Schouten HJ, Koudstaal J, Kwee WS, Willebrand D, Sturmans F, Arends JW. Sources of interobserver variation in histopathological grading of cervical dysplasia. J Clin Epidemiol. 1992;45:785-90.
10. Grenko RT, Abendroth CS, Frauenhoffer EE, Ruggiero FM, Zaino RJ. Variance in the interpretation of cervical biopsy specimens obtained for atypical squamous cells of undetermined significance. Am J Clin Pathol. 2000;114:735-40.
11. Klaes R, Benner A, Friedrich T, Ridder R, Herrington S, Jenkins D, Kurman RJ, Schmidt D, Stoler M, von Knebel Doeberitz M. p16INK4a immunohistochemistry improves interobserver agreement in the diagnosis of cervical intraepithelial neoplasia. Am J Surg Pathol. 2002;26:1389-99.
12. Stoler MH, Rhodes CR, Whitbeck A, Wolinsky SM, Chow LT, Broker TR. Human papillomavirus type 16 and 18 gene expression in cervical neoplasias. Hum Pathol. 1992;23:117-28.
13. Woodhouse SL, Stastny JF, Styer PE, Kennedy M, Praestgaard AH, Davey DD. Interobserver variability in subclassification of squamous intraepithelial lesions: Results of the College of American Pathologists Interlaboratory Comparison Program in Cervicovaginal Cytology. Arch Pathol Lab Med. 1999;123:1079-84.
14. Ismail SM, Colclough AB, Dinnen JS, Eakins D, Evans DM, Gradwell E, O'Sullivan JP, Summerell JM, Newcombe RG. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. BMJ. 1989;298:707-10.
15. Robertson AJ, Anderson JM, Beck JS, Burnett RA, Howatson SR, Lee FD, Lessells AM, McLaren KM, Moss SM, Simpson JG. Observer variability in histopathological reporting of cervical biopsy specimens. J Clin Pathol. 1989;42:231-8.
16. Cocker J, Fox H, Langley FA. Consistency in the histological diagnosis of epithelial abnormalities of the cervix uteri. J Clin Pathol. 1968;21:67-70.
17. Horn LC, Reichert A, Oster A, Arndal SF, Trunk MJ, Ridder R, Rassmussen OF, Bjelkenkrantz K, Christiansen P, Eck M, Lorey T, Skovlund VR, Ruediger T, Schneider V, Schmidt D. Immunostaining for p16INK4a used as a conjunctive tool improves interobserver agreement of the histologic diagnosis of cervical intraepithelial neoplasia. Am J Surg Pathol. 2008;32:502-12.
18. Sayed K, Korourian S, Ellison DA, Kozlowski K, Talley L, Horn HV, Simpson P, Parham DM. Diagnosing cervical biopsies in adolescents: The use of p16 immunohistochemistry to improve reliability and reproducibility. J Low Genit Tract Dis. 2007;11:141-6.
19. Zhang Q, Kuhn L, Denny LA, De Souza M, Taylor S, Wright TC. Impact of utilizing p16INK4A immunohistochemistry on estimated performance of three cervical cancer screening tests. International Journal of Cancer. 2007;120:351-6.
20. Dijkstra MG, Heideman DA, de Roy SC, Rozendaal L, Berkhof J, van Krimpen K, van Groningen K, Snijders PJ, Meijer CJ, van Kemenade FJ. p16(INK4a) immunostaining as an alternative to histology review for reliable grading of cervical intraepithelial lesions. J Clin Pathol. 2010;63:972-7.
21. Castle PE, Stoler MH, Solomon D, Schiffmanx M, Group ftA. The Relationship of Community Biopsy-Diagnosed Cervical Intraepithelial Neoplasia Grade 2 to the Quality Control Pathology-Reviewed Diagnoses: An ALTS Report. American Journal of Clinical Pathology. 2007;127:805-15.
22. Carreon JD, Sherman ME, Guillén D, Solomon D, Herrero R, Jerónimo J, Wacholder S, Rodríguez AC, Morales J, Hutchinson M, Burk RD, Schiffman M. 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.
23. Parker MF, Zahn CM, Vogel KM, Olsen CH, Miyazawa K, O'Connor DM. Discrepancy in the interpretation of cervical histology by gynecologic pathologists. Obstet Gynecol. 2002;100:277-80.

24. Vinyuvat S, Karalak A, Suthipintawong C, Tungsinmunkong K, Kleebkaow P, Trivijitsilp P, Siriaunkgul S, Triratanachat S, Khunamornpong S, Chuangsuwanich T, Settakorn J. Interobserver reproducibility in determining p16 overexpression in cervical lesions: Use of a combined scoring method. Asian Pac J Cancer Prev. 2008;9:653-7.
25. Gurrola-Diaz CM, Suarez-Rincon AE, Vazquez-Camacho G, Buonocunto-Vazquez G, Rosales-Quintana S, Wentzensen N, von Knebel Doeberitz M. P16INK4a immunohistochemistry improves the reproducibility of the histological diagnosis of cervical intraepithelial neoplasia in cone biopsies. Gynecol Oncol. 2008;111:120-4.
26. Bergeron C, Ordi J, Schmidt D, Trunk MJ, Keller T, Ridder R. Conjunctive p16INK4a testing significantly increases accuracy in diagnosing high-grade cervical intraepithelial neoplasia. Am J Clin Pathol. 2010;133:395-406.
27. Yildiz IZ, Usubutun A, Firat P, Ayhan A, Kucukali T. Efficiency of immunohistochemistry p16 expression and HPV typing in cervical squamous intraepithelial lesion grading and review of the p16 literature. Pathol Res Pract. 2007;203:445-9.
28. Crum CP. Our wages of CIN. Obstet Gynecol. 2012;120:1261-2.
29. Massad LS, Einstein MH, Huh WK, Katki HA, Kinney WK, Schiffman M, Solomon D, Wentzensen N, Lawson HW; 2012 ASCCP Consensus Guidelines Conference. 2012 updated consensus guidelines for the management of abnormal cervical cancer screening tests and cancer precursors. J Low Genit Tract Dis. 2013;17:S1-S27.
30. Usubutun A, Mutter GL, Saglam A, Dolgun A, Ozkan EA, Ince T, Akyol A, Bulbul HD, Calay Z, Eren F, Gumurdulu D, Haberal AN, Ilvan S, Karaveli S, Koyuncuoglu M, Muezzinoglu B, Muftuoglu KH, Ozdemir N, Ozen O, Baykara S, Pestereli E, Ulukus EC, Zekioglu O. Reproducibility of endometrial intraepithelial neoplasia diagnosis is good, but influenced by the diagnostic style of pathologists. Mod Pathol. 2012;25:877-84.