JPTM

Interobserver Reproducibility of PD-L1 Biomarker in Non-small Cell Lung Cancer: A Multi-Institutional Study by 27 Pathologists

Sunhee Chang*, Hyung Kyu Park[1]*, Yoon-La Choi[2], Se Jin Jang[3],
Cardiopulmonary Pathology Study Group of the Korean Society of Pathologists

Department of Pathology, Inje University Ilsan Paik Hospital, Goyang; [1]Department of Pathology, Konkuk University Medical Center, Konkuk University School of Medicine, Seoul; [2]Department of Pathology and Translational Genomics, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul; [3]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Background: Assessment of programmed cell death-ligand 1 (PD-L1) immunohistochemical staining is used for treatment decisions in non-small cell lung cancer (NSCLC) regarding use of PD-L1/programmed cell death protein 1 (PD-1) immunotherapy. The reliability of the PD-L1 22C3 pharmDx assay is critical in guiding clinical practice. The Cardiopulmonary Pathology Study Group of the Korean Society of Pathologists investigated the interobserver reproducibility of PD-L1 staining with 22C3 pharmDx in NSCLC samples. **Methods:** Twenty-seven pathologists individually assessed the tumor proportion score (TPS) for 107 NSCLC samples. Each case was divided into three levels based on TPS: <1%, 1%–49%, and ≥50%. **Results:** The intraclass correlation coefficient for TPS was 0.902 ± 0.058. Weighted κ coefficient for 3-step assessment was 0.748 ± 0.093. The κ coefficients for 1% and 50% cut-offs were 0.633 and 0.834, respectively. There was a significant association between interobserver reproducibility and experience (formal PD-L1 training, more experience for PD-L1 assessment, and longer practice duration on surgical pathology), histologic subtype, and specimen type. **Conclusions:** Our results indicate that PD-L1 immunohistochemical staining provides a reproducible basis for decisions on anti–PD-1 therapy in NSCLC.

Key Words: Programmed cell death-ligand 1; Reproducibility; Observer variation; Immunohistochemistry

Immunotherapies with checkpoint inhibitor programmed death-ligand 1 (PD-L1)/programmed cell death protein 1 (PD-1) antibodies have shown encouraging results in patients with advanced non-small cell lung cancer (NSCLC).[1-4] Assessment of PD-L1 immunohistochemical staining has been developed for pathology laboratories to aid in selecting patients who will benefit from PD-1/PD-L1–targeted therapy.[1,5] PD-L1 immunohistochemistry is currently the most useful biomarker because of the wide availability of formalin-fixed, paraffin-embedded tissues, the relatively low cost, and widespread use in pathology laboratories, particularly in contrast to molecular pathology-based methods.[1,6]

Pembrolizumab (Keytruda, Merck & Co., Inc., Kenilworth, NJ, USA), approved by the U.S. Food and Drug Administration (FDA) and Korea Ministry of Food and Drug Safety (MFDS), is a first-line therapy PD-1 inhibitor for patients with advanced NSCLC.[1] Dako PD-L1 immunohistochemical (IHC) 22C3 pharmDx was used to determine PD-L1 expression in patients with advanced NSCLC during the clinical phase 1 trial (KEYNOTE-001).[2,7] The PD-L1 IHC 22C3 pharmDx assay is the first companion diagnostic assay for PD-L1 approved by the FDA and MFDS.[1,5] Currently, four PD-L1 assays using four different PD-L1 antibodies (22C3, 28-8, SP263, and SP142) on two different IHC platforms (Dako and Ventana) are approved by the FDA and Korea Food & Drug Administration. Each assay has its own scoring system. In the era of immunotherapy, reliability of assay results is critical to predict the likelihood of response to anti–PD-1 or anti–PD-L1 therapy. A few studies have investigated the reproducibility of assessing PD-L1 tests in NSCLC tissue samples.[8-10]

This study aimed to investigate the interobserver reproducibility of assessing PD-L1 expression in NSCLC tissue samples. Furthermore, association with observer factors and reproducibility of PD-L1 assessment was also investigated.

MATERIALS AND METHODS

Study material

A total of 107 cases of NSCLC were selected from the archives of the Department of Pathology of Samsung Medical Center from October 2016 and December 2016. Of these, 22 tissue samples were from resections, 66 tissue samples were from computed tomography-guided core biopsy, and 19 tissue samples were cell blocks of endobronchial ultrasound-guided transbronchial needle aspirate (EBUS-TBNA). The study material comprised 66 adenocarcinomas, 33 squamous cell carcinomas, and eight other non-small cell lung cancers. Selected tumor samples contained more than 100 cells per sample.

Immunohistochemical staining and evaluation

Tissue samples were stained for PD-L1 with the 22C3 phar-mDx Kit (Agilent Technologies, Santa Clara, CA, USA) on the Dako Autostainer Link 48 platform (Agilent Technologies). Deparaffinization, rehydration, and target retrieval procedures were performed using EnVision FLEX Target Retrieval solution (1×, low pH) and EnVision FLEX wash buffer (1×). The tissue samples were then placed on the Autostainer Link 48. This instrument performed the staining process by applying appropriate reagent, monitoring incubation time, and rinsing slides between reagents. The reagent times were preprogrammed in the Dako Link software. A sample with primary antibody omitted was used as a negative control. Samples were subsequently counterstained with hematoxylin and mounted in non-aqueous, permanent mounting media. The stained slides were scanned by a Aperio scanner (Leica Biosystems, Buffalo Grove, IL, USA). Pathologists scored the virtual images using ImageScope software (Leica Biosystems).

Tumor proportion score (TPS) was defined as the percentage of viable tumor cells with any perceptible membrane staining irrespective of staining intensity. Normal cells and tumor-associated immune cells were excluded from scoring. Each case was divided into three levels based on TPS: <1% (no PD-L1 expression), 1%–49% (PD-L1 expression), or ≥50% (high PD-L1 expression) (Fig. 1).
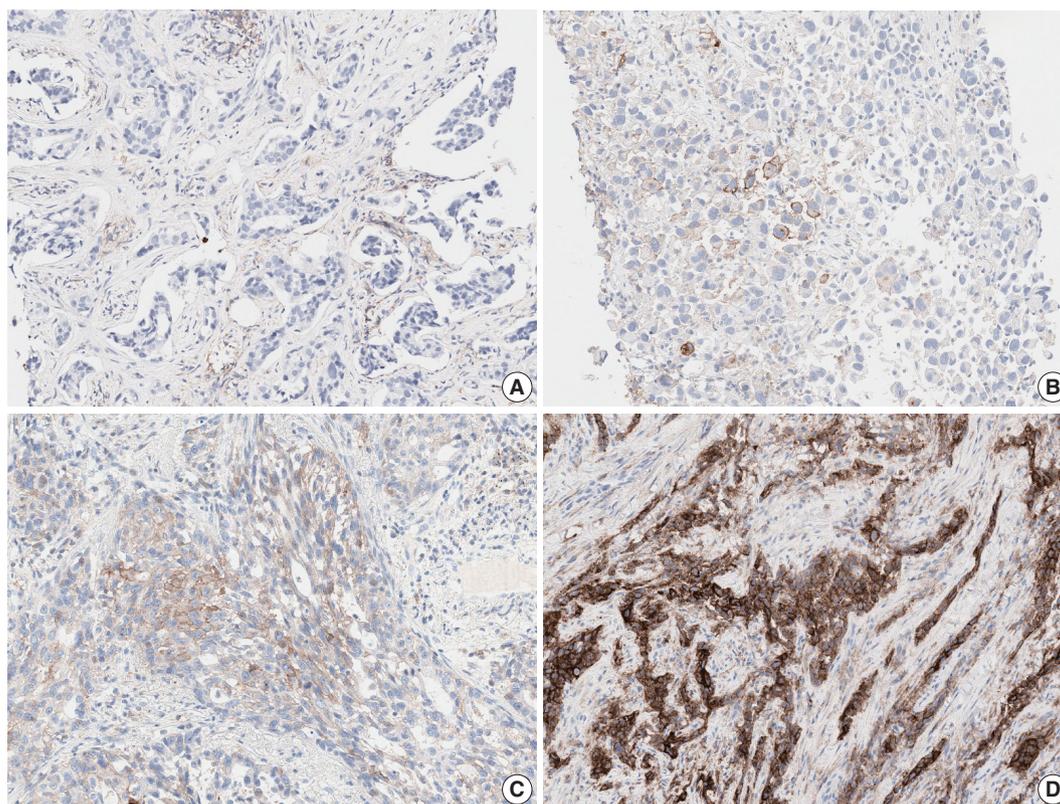
Fig. 1. Programmed cell death-ligand 1 (PD-L1) immunohistochemistry results in non-small cell lung cancer patients using 22C3 antibody on fully automated Dako Autostainer Link 48 platform. (A) Negative staining for PD-L1. (B) PD-L1 tumor proportion score (TPS) of 10%. (C) PD-L1 TPS of 70%. (D) PD-L1 TPS of 100%.

Participating pathologists

All slides were independently evaluated by 20 pulmonary pathology specialists and seven surgical pathology fellows. Eleven experts participated in a 1-day training course by Dako (Agilent Technologies, Santa Clara, CA, USA) relating to evaluation of the PD-L1 22C3 assay. Participant practice duration ranged from 1 to 37 years with a median of 13 years. Sixteen of the 27 pathologists gained experience in PD-L1 assessment by practicing daily with two to 500 cases per day for 2 months before starting this study, and four of the 16 pathologists had assessed more than 100 cases.

Statistical analysis

The gold standard PD-L1 TPS was established as TPS and assessed by highly trained and experienced experts. To assess the concordance of TPS between the 27 pathologists and the gold standard, intraclass correlation coefficients (ICCs) were calculated. The weighted kappa (κ) coefficient was calculated to evaluate concordance of the 3-step assessment between pathologists and the gold standard. Agreement for 1% and 50% cut-offs was assessed using overall percent agreement (OPA), positive and negative percent agreement, and Cohen's κ coefficient. Correlations between TPS and experience for PD-L1 test experience or practice duration were investigated using Spearman's rank correlation coefficient. Correlations between the 3-step assessment and PD-L1 test experience or practice duration were also investigated using Spearman's rank correlation coefficient. Differences in the concordance of TPS, 3-step assessment, and 1% and 50% cut-offs were compared between the expert group and fellow group using Wilcoxon rank sum tests. Correlations between TPS or 3-step assessment and training were assessed using Wilcoxon rank sum test. Correlations between TPS or 3-step assessment and histologic subtype or specimen type were assessed using Kruskal-Wallis Test.

An ICC is interpreted as follows: below 0.3 indicates poor agreement, 0.5 indicates moderate agreement, 0.7 indicates strong agreement, and 0.85 or more indicates almost perfect agreement. A κ coefficient of 0.4 or less is poor to fair agreement, greater than 0.4 to 0.6 is moderate agreement, greater than 0.6 to 0.8

is substantial agreement, and greater than 0.8 is almost perfect agreement. Spearman's rank correlation coefficient is interpreted as follows: below 0.3 indicates little if any (linear) correlation, between 0.3 and 0.5 indicates low correlation, between 0.5 and 0.7 indicates moderate correlation, between 0.7 and 0.9 indicates high correlation, and 0.9 or more indicates very high correlation. A p-value of < .05 was considered statistically significant. Statistical analyses were performed using SAS ver. 9.4 (SAS Institute Inc., Cary, NC, USA).

Ethics statement

This study was approved by the Institutional Review Board of Inje University Ilsan Paik Hospital with a waiver of informed consent (2018-08-009).

RESULTS

Interobserver reproducibility

Among 107 samples, 22 samples (20.6%) had a TPS <1%, 40 samples (37.4%) had a TPS between 1% and 49%, and 45 samples (42.1%) had a TPS ≥50%. The ICC for TPS was 0.902 ± 0.058, indicating almost perfect agreement (Table 1). Weighted κ coefficient for the 3-step assessment was 0.748 ± 0.093, indicating substantial agreement. Using ≥1% stained tumor cells as the cut-off for a positive test, the κ coefficient was calculated as 0.633 ± 0.111, and OPA was 86.2 ± 5.5% (Table 2). The κ coefficient was 0.834 ± 0.095, and the OPA was 92.1 ± 4.4% for cut-off ≥50%. These results indicate substantial agreement for 1% and almost perfect agreement for 50% cut-offs. The interobserver reproducibility was greater at the 50% cut-off than at the 1% cut-off.

Factors influencing interobserver reproducibility

There was significant association between interobserver reproducibility and specimen type (Table 3). The ICC for TPS was significantly higher in the resection group (0.926 ± 0.050) than in the EBUS-TBNA group (0.887 ± 0.099). The κ coefficient for the 3-step assessment was significantly higher in the biopsy group (0.776 ± 0.104) than in the EBUS-TBNA group (0.669 ± 0.089).

Table 1. Intraclass correlation coefficient of the tumor proportion score

	ICC for TPS	p-value	Weighted Kappa for 3 tiered assessment	p-value
Total (n=26)	0.902 ± 0.058		0.748 ± 0.093	
Expert (n=19)	0.919 ± 0.034	.037	0.772 ± 0.073	.043
Fellow (n=7)	0.859 ± 0.088		0.680 ± 0.115	

ICC, Intraclass correlation coefficient; TPS, tumor proportion score.

The EBUS-TBNA group showed almost perfect agreement for the 50% cut-off, but fair agreement for the 1% cut-off.

ICCs for TPS were influenced by histologic subtype (Table 4). The squamous cell carcinoma group showed higher agreement than the adenocarcinoma group. There was no significant difference between κ coefficients for the 3-step assessment.

There was significant association between interobserver reproducibility and experience (formal PD-L1 training, more experience for PD-L1 test, and longer practice duration on surgical pathology). The ICC for TPS was significantly higher in the trained group (0.922 ± 0.034) than in the untrained group (0.875 ± 0.074) (Table 5). The κ coefficient for the 3-step assessment was also significantly higher in the trained group (0.776 ± 0.079) than in the untrained group (0.709 ± 0.101). There was low linear correlation between ICC for TPS and experience for PD-L1 assessment (Spearman correlation coefficient = 0.422) (Table 6). There was no correlation between the κ coefficient for the 3-step assessment and experience for PD-L1 assessment. The ICC for TPS was significantly higher in the expert group (91.9% ± 3.4%) than in the fellow group (85.9% ± 8.8%) (Table 1). The κ coefficient for the 3-step assessment was also significantly higher in the expert

group (0.772 ± 0.073) than in the fellow group (0.680 ± 0.115). The κ coefficient for the 3-step assessment was significantly different in biopsy specimens between the expert group (0.807 ± 0.034) and the fellow group (0.693 ± 0.129) (p = 0.0013). In EBUS-TBNA and resection specimens, the κ coefficient of the expert group was slightly higher than that of the fellow group but there was no statistically significant difference between the two groups. There were no differences in κ coefficients at 1% and 50% cut-offs between the expert group and fellow group (Table 2).

DISCUSSION

We investigated the interobserver reproducibility of assessment of PD-L1 expression in NSCLC and observed good interobserver agreement for PD-L1 scoring. Pathologists were highly concordant for TPS with an ICC of 0.902. Rimm et al.[8] examined the interobserver reproducibility for 22C3, 28-8, SP142, and E1L3N PD-L1 assays. Ninety samples were assessed by 13 pathologists. ICC was 0.882 for the 22C3 assay. The ICCs for 28-8, SP142, and E1L3N assays were 0.832, 0.869, and 0.859, respectively, showing high concordance. The findings suggest that PD-L1 assay is a reliable method for assessing PD-L1 expression in tumor cells.

We report good interobserver agreement for 1% and 50% cut-offs with κ coefficients of 0.633 and 0.834, respectively (Table 7). Cooper et al.[9] investigated the interobserver reproducibility for

Table 2. Interobserver reproducibility of the cut-off

	1% Cut-off	p-value	50% Cut-off	p-value
Cohen's κ coefficient				
Total (n = 26)	0.633 ± 0.111		0.834 ± 0.095	
Expert (n = 19)	0.656 ± 0.104	.068	0.858 ± 0.072	.082
Fellow (n = 7)	0.570 ± 0.113		0.768 ± 0.123	
OPA (%)				
Total (n = 26)	86.2 ± 5.5		92.1 ± 4.4	
Expert (n = 19)	87.3 ± 4.6	.067	93.2 ± 3.3	.075
Fellow (n = 7)	83.2 ± 6.8		89.1 ± 5.5	
NPA (%)				
Total (n = 26)	85.7 ± 16.0		95.4 ± 4.3	
Expert (n = 19)	86.8 ± 17.1	.150	95.3 ± 4.5	.884
Fellow (n = 7)	82.5 ± 13.5		95.4 ± 4.0	
PPA (%)				
Total (n = 26)	86.3 ± 8.4		87.5 ± 12.0	
Expert (n = 19)	87.4 ± 7.6	.385	9.2 ± 9.2	.149
Fellow (n = 7)	83.4 ± 1.2		8.3 ± 16.4	

OPA, overall percent agreement; NPA, negative percent agreement; PPA, positive percent agreement.

Table 4. Impact of histologic subtype on interobserver reproducibility

	SCC (n = 33)	ADC (n = 66)	p-value
ICC for TPS	0.877 ± 0.071	0.917 ± 0.053	.024
κ for 3-step evaluation	0.757 ± 0.107	0.753 ± 0.089	.073

SCC, squamous cell carcinoma; ADC, adenocarcinoma; ICC, intraclass correlation coefficient; TPS, tumor proportion score.

Table 5. Impact of training on interobserver reproducibility

	Training		p-value
	Yes (n = 15)	No (n = 11)	
ICC for TPS	0.922 ± 0.034	0.875 ± 0.074	.043
κ for 3-step assessment	0.776 ± 0.079	0.709 ± 0.101	.026

ICC, intraclass correlation coefficient; TPS, tumor proportion score.

Table 3. Impact of specimen type on interobserver reproducibility

	EBUS-TBNA (n = 19)	Biopsy (n = 66)	Resection (n = 22)	p-value
ICC for TPS	0.887 ± 0.099	0.899 ± 0.061	0.926 ± 0.050	.023
κ for 3-step evaluation	0.669 ± 0.089	0.776 ± 0.104	0.716 ± 0.126	.001
κ for 1% cutoff	0.383 ± 0.134	0.713 ± 0.137	0.538 ± 0.209	
κ for 50% cutoff	0.832 ± 0.128	0.830 ± 0.097	0.839 ± 0.118	

EBUS-TBNA, Endobronchial ultrasound-guided transbronchial needle aspiration; ICC, intraclass correlation coefficient; TPS, tumor proportion score.

Table 6. Impact of experience on interobserver reproducibility

	ICC for TPS		κ coefficient for 3-step evaluation	
	Spearman correlation	p-value	Spearman correlation	p-value
PD-L1 test experience	0.422	.032	0.277	.170
Practice duration	0.477	.014	0.527	.005

ICC, intraclass correlation coefficient; TPS, tumor proportion score; PD-L1, programmed cell death-ligand 1.

Table 7. Summary of the interobserver reproducibility study

	Sample	Observer	1% Cut-off	50% Cut-off
Rimm et al.[8]	90	13	κ=0.537[a]	κ=0.749[a]
Cooper et al.[9]	108	10	κ=0.68	κ=0.58
Brunnström et al.[10]	55	7	2%–20%[b]	0%–2%[b]
Current study	107	27	κ=0.633	κ=0.834

[a]For the mean of all four assay (22C3, 28-8, SP142, and E1L3N); [b]Differently classified cases by any one pathologist compared with consensus.

assessment of the 22C3 PD-L1 assay. Two separate sample sets of 60 samples each were designed for the 1% and 50% cut-offs that contained equally distributed PD-L1 positive and negative samples. The sample set for the 1% cut-off contained 10 positive samples close to the cut-off, and the sample set for the 50% cut-off contained 20 negative or positive samples close to the cut-off. Ten pathologists assessed a sample set of 108 samples obtained after pooling the 1% cut-off and the 50% cut-off sample sets together. The κ coefficient was 0.68 for the 1% cut-off and 0.58 for the 50% cut-off. Brunnström et al. examined interobserver reproducibility for the 28-8, 22C3, SP142, and SP263 assays.[10] Seven pathologists assessed 55 samples. For the 22C3 assay, 2%–20% of cases were differently classified by any one pathologist compared to the consensus at the 1% cut-off, and 0%–2% of the cases were differently classified by any one pathologist compared to the consensus at the 50% cut-off. For all four assays, there were 0%–20% and 0%–5% differently classified cases at 1% and 50% cut-offs, respectively. Variation in the number of differently classified cases by any one pathologist compared to the consensus was statistically significant between cut-offs. The number of differently classified cases was significantly lower for the SP142 assay compared to that for the other three assays. This difference was probably because there were many obviously negative cases for SP142.[10] Rimm et al.[8] reported that κ coefficients for the mean of all 4 assays were 0.537 and 0.749 at 1% and 50% cut-offs, respectively. Interobserver concordance for the PD-L1 assay was higher at the 50% cut-off than at the 1% cut-off, except for the results of Cooper et al.[9] It is possible that Cooper et al.[9] used sample sets to artificially enrich with samples close to the cut-offs, which is not possible in clinical practice.

The interobserver concordance for TPS and the 3-step assess-ment correlated with practice duration and was higher in the expert group than in the fellow group. This is probably because pulmonary pathologists were familiar with the varied morphologies of cancer and cancer-associated immune cells and had experience assessing other immunohistochemistry biomarkers, such as anaplastic lymphoma kinase, human epidermal growth factor receptor 2, and estrogen receptor.[9] There was no difference in concordance between the expert group and fellow group for 1% and 50% cut-offs, similar to the results of Brunnström et al.[10] Experience in conducting the PD-L1 test impacted interobserver concordance for TPS but not for 3-step assessment. These analyses suggested that the currently used 1% and 50% cut-offs are relatively reliable regardless of pathologist experience.

Our study reported that 1-day training improved interobserver reproducibility, whereas 1-hour training had no or very little impact on interobserver reproducibility.[9] Therefore, 1-day training may be more effective than 1-hour training.

Tumor histologic subtype and specimen type influenced interobserver reproducibility, which was not previously reported. EBUS-TBNA showed lower interobserver agreement than resection and biopsy specimens at the 1% cut-off. Although squamous cell carcinoma showed higher agreement for TPS than adenocarcinoma, no significant difference was observed for the 3-step assessment. Strong membranous staining of macrophages, nonspecific cytoplasmic staining of tumor cells, weak and/or partial membranous staining of tumors cells, heterogeneous staining intensities, and patchy staining are well-known interpretation pitfalls in assessing any PD-L1 assay (Fig. 2).[8-10] While using the 1% cut-off, misinterpretation of very few or even single cells may lead to false positive or false negative results. Training and an external quality assessment program should be organized with special focus on difficult cases and on assessing the 1% cut-off. Guidelines including examples and strategies for difficult cases should be developed.

The limitations of our study include the lack of a "true gold standard" and outcome data for therapy. However, gold standard assessment was undertaken by highly trained experienced specialists. The following were strengths of the present study: it is more representative of real clinical practice than previous studies,
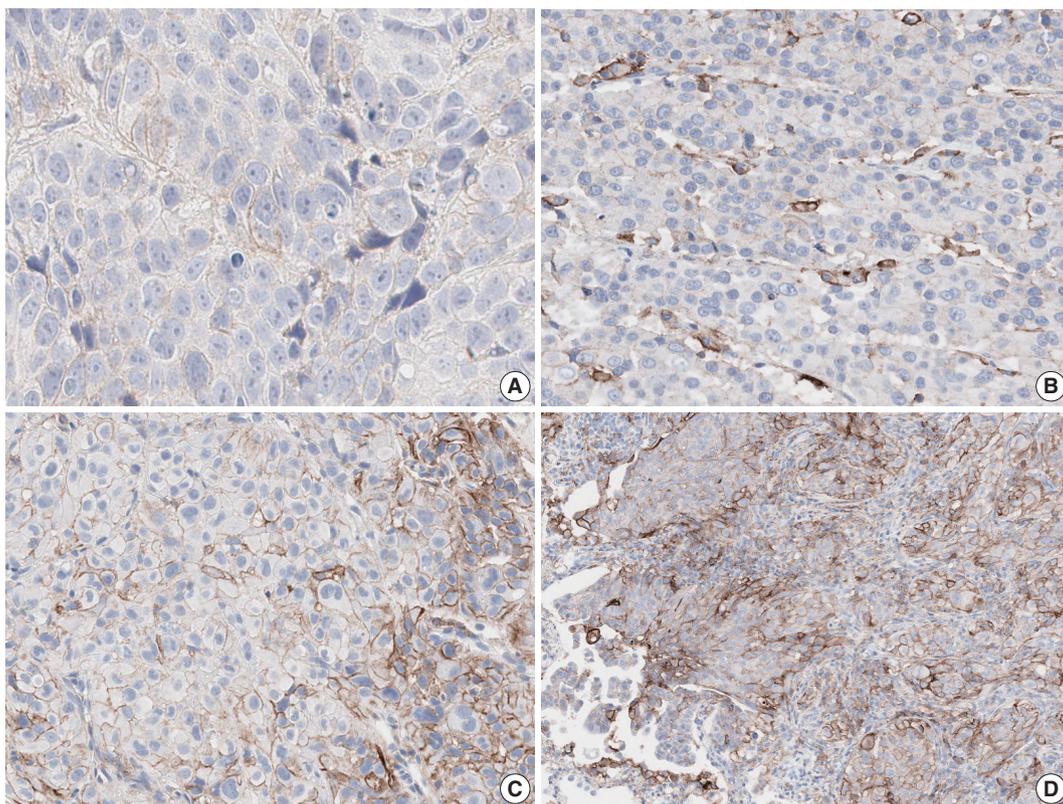
Fig. 2. (A) Few tumor cells show weak and partial membrane staining for programmed cell death-ligand 1 (PD-L1) antibody. (B) Tumor associated immune cells show strong staining with lack of PD-L1 staining in tumor cells. (C) Tumor cells show heterogeneous membrane staining pattern with various staining intensities. (D) Tumor shows patchy membrane staining pattern.

whole sections from many samples were used to evaluate the reproducibility of PD-L1 assays, more observers than previous studies, and participating pathologists worked in different centers and had different levels of experience.

In conclusion, our results indicate that PD-L1 staining provides a reliable basis for decisions regarding anti-PD-1 therapy in NSCLC. Although interobserver agreement for the 1% cut-off was relatively lower, it was substantial and acceptable. Better training, longer assay experience, and an external quality assessment program could improve interobserver reproducibility for the 1% cut-off.

ORCID

Sunhee Chang: https://orcid.org/0000-0002-7775-4711
Hyung Kyu Park: https://orcid.org/0000-0002-5972-3516
Yoon-La Choi: http://orcid.org/0000-0002-5788-5140
Se Jin Jang: http://orcid.org/0000-0001-8239-4362

Author Contributions

Conceptualization: YLC.
Data curation: SC, HKP, YLC.
Formal analysis: HKP, YLC.
Funding acquisition: SC, SJJ.
Investigation: SC, HKP, YLC.
Methodology: SC, HKP, YLC.
Project administration: YLC.
Resources: YLC.
Supervision: YLC, SJJ.
Validation: SC, HKP, YLC.
Visualization: SC, HKP, YLC.
Writing—original draft: SC.
Writing—review & editing: SC, YLC, SJJ.

Conflicts of Interest

The authors declare that they have no potential conflicts of interest.

REFERENCES

1. Guan J, Lim KS, Mekhail T, Chang CC. Programmed death ligand-1 (PD-L1) expression in the programmed death receptor-1 (PD-1)/PD-L1 blockade: a key player against various cancers. Arch Pathol Lab Med 2017; 141: 851-61.

2. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N Engl J Med 2016; 375: 1823-33.

3. Langer CJ, Gadgeel SM, Borghaei H, et al. Carboplatin and pemetrexed with or without pembrolizumab for advanced, non-squamous non-small-cell lung cancer: a randomised, phase 2 cohort of the open-label KEYNOTE-021 study. Lancet Oncol 2016; 17: 1497-508.

4. Gettinger S, Rizvi NA, Chow LQ, et al. Nivolumab monotherapy for first-line treatment of advanced non-small-cell lung cancer. J Clin Oncol 2016; 34: 2980-7.

5. Roach C, Zhang N, Corigliano E, et al. Development of a companion diagnostic PD-L1 immunohistochemistry assay for pembrolizumab therapy in non-small-cell lung cancer. Appl Immunohistochem Mol Morphol 2016; 24: 392-7.

6. Cree IA, Booton R, Cane P, et al. PD-L1 testing for lung cancer in the UK: recognizing the challenges for implementation. Histopathology 2016; 69: 177-86.

7. Herbst RS, Baas P, Kim DW, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. Lancet 2016; 387: 1540-50.

8. Rimm DL, Han G, Taube JM, et al. A prospective, multi-institutional, pathologist-based assessment of 4 immunohistochemistry assays for PD-L1 expression in non-small cell lung cancer. JAMA Oncol 2017; 3: 1051-8.

9. Cooper WA, Russell PA, Cherian M, et al. Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non-small cell lung cancer. Clin Cancer Res 2017; 23: 4569-77.

10. Brunnström H, Johansson A, Westbom-Fremer S, et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: interpathologist variability is higher than assay variability. Mod Pathol 2017; 30: 1411-21.