ORIGINAL ARTICLE

The Reproducibility of Histopathologic Assessments of Programmed Cell Death-Ligand 1 Using Companion Diagnostics in NSCLC

Pei Yuan, MD, Changyuan Guo, MD, PhD, Lin Li, MD, PhD, Lei Guo, BS, Fanshuang Zhang, PhD, Jianming Ying, MD, PhD*

Department of Pathology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, People's Republic of China

Received 20 May 2020; revised 22 July 2020; accepted 22 September 2020
Available online - 5 October 2020

ABSTRACT

Introduction: Accurate results on the status of programmed cell death-ligand 1 (PD-L1) rely on not only the quality of immunohistochemistry testing but also the accuracy of the pathologic assessments. We explored the intraobserver and interobserver reproducibility of the interpretations for the companion diagnostics, the Dako PD-L1 22C3 pharmDx kit (Dako North America, Inc, Carpinteria, CA) and the VENTANA PD-L1 (SP263, Ventana Medical Systems, Inc, Tucson, AZ) assay, and the consistency between microscopic and digital interpretations of PD-L1.

Methods: A total of 150 surgical specimens diagnosed as NSCLC from December 2013 to July 2017 were included in this study. Twenty pathologists from different medical centers were enrolled to interpret the results of PD-L1 on the same day. A total of 100 sections were stained with the 22C3 clone and scored for the interobserver reproducibility, 20 cases of which were interpreted twice to assess the intraobserver reproducibility, and 50 cases of which were scanned into digital images to measure the consistency between microscopic and digital interpretations. A total of 44 sections were stained with the SP263 clone and scored for the interobserver reproducibility.

Results: For the intraobserver reproducibility of 22C3, the overall percent agreements were 92.0% and 89.0% for binary tumor evaluation at the cutoffs of 1% and 50%, respectively. The reliability among the pathologists revealed a substantial agreement for 22C3, whereas it revealed a substantial agreement at the cutoff of 1% and moderate agreement at the cutoffs of 25% and 50% for SP263. Microscopic and digital interpretations of PD-L1 revealed good consistency.

Conclusions: Intraobserver and interobserver reproducibility of the interpretations for PD-L1 was high using the 22C3 clone but lower for the SP263 clone. Corresponding training on such assessments, especially on the cases around the specific cutoffs, is essential for markedly improving such reproducibility. Digital imaging could improve the reproducibility of interpretation for PD-L1 among pathologists.

Keywords: Assessment; PD-L1; 22C3; SP263; Reproducibility

Introduction

Lung cancer is the most common cause of cancer death worldwide.[1] NSCLC accounts for 85% of lung cancer cases and is often diagnosed at a late stage; by this stage, the opportunity to undergo surgery has

*Corresponding author.

Disclosure: The authors declare no conflict of interest.

Address for correspondence: Jianming Ying, MD, PhD, Department of Pathology, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences, No. 17, Panjiayuan Nanli, Chaoyang District, 100021 Beijing, People's Republic of China. E-mail: jmying@hotmail.com

Cite this article as: Yuan P, et al. The Reproducibility of Histopathologic Assessments of Programmed Cell Death-Ligand 1 Using Companion Diagnostics in NSCLC. JTO Clin Res Rep 2:100102

ISSN: 2666-3643

https://doi.org/10.1016/j.jtocrr.2020.100102

already passed for many patients. Several large clinical studies have revealed the benefits of immunotherapy for advanced NSCLC; there have been particularly promising breakthroughs with immune checkpoint inhibitors for NSCLC.[2-6] On the basis of the results of the CheckMate, KEYNOTE, OAK, and PACIFIC trials, the U.S. Food and Drug Administration (FDA) has approved pembrolizumab, nivolumab, atezolizumab, and durvalumab for NSCLC[7-10] and four auxiliary diagnostic kits (Dako Programmed Cell Death-Ligand 1 [PD-L1] immunohistochemistry [IHC] 22C3 and 28-8 pharmDx assays [Dako North America, Inc, Carpinteria, CA] and VENTANA PD-L1 SP263 and SP142 assays [Ventana Medical Systems, Inc, Tucson, AZ]). The Dako PD-L1 IHC 22C3 pharmDx has been approved by the FDA as a companion diagnostic for use with pembrolizumab in NSCLC using 1% and 50% as the cutoffs. The Dako 28-8 pharmDx and VENTANA SP142 kits were also approved as complementary diagnostics for use with nivolumab and atezolizumab, respectively. The VENTANA SP263 assay served as a complementary diagnostic for use with durvalumab in NSCLC as approved by the FDA. In addition, it is approved as a companion diagnostic for use with durvalumab and pembrolizumab and as a complementary diagnostic for use with nivolumab by Conformité Européenne with different cutoffs on the basis of the results of an AstraZeneca comparison study.[11]

Accurate results on the status of PD-L1 rely on both the quality of IHC testing and the accuracy of pathologic assessments. Several studies have explored the concordance among different PD-L1 antibody clones and have revealed that 22C3, 28-8, and SP263 have good staining consistency for tumor cells but poor consistency for immune cells.[11-15] The Blueprint PD-L1 Immunohistochemistry Comparability Project indicated that, although these three assays had similar analytical performance for PD-L1 expression, the interchanging of assays and cutoffs would lead to misclassification of the PD-L1 status in some patients. A few studies explored both assay compatibility and consistency of pathologists' assessments and uniformly revealed that interpathologist variability was higher than assay variability.[11,12,16,17] Therefore, it seems that, when using the approved assays, a major challenge could be the variability of pathologists' assessments. Several studies have explored the interobserver and intraobserver reproducibility of such assessments; however, these studies were limited by too few trained pathologists, a small sample size, or the use of just a single antibody,[18-20] making it easy to conclude that there was high reproducibility.

Thus, in our study, we aimed to include a greater number of samples and pathologists to explore the following: (1) the intraobserver and interobserver reproducibility regarding the interpretation of the 22C3 clone at the cutoffs of 1% and 50%; (2) the interobserver reproducibility regarding the interpretation of the SP263 clone at the cutoffs of 1%, 25%, and 50%; (3) the consistency between microscopic and digital interpretations; and (4) the influences of professional titles, specialty, and the number of working years on the consistency of interpretation.

Materials and Methods

Case Selection

A total of 150 surgical specimens diagnosed as NSCLC were randomly enrolled from December 2013 to July 2017 at the Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, People's Republic of China. Considering the retrospective nature of the design, this study was approved with no additional patient consent required.

PD-L1 IHC Assays and Slide Scanning

A total of 150 paraffin blocks were continuously sliced until at least three tissue sections were obtained with no less than 100 tumor cells identified on the hematoxylin and eosin-stained sections. These sections were then stained for PD-L1. A total of 100 cases were stained with the Dako PD-L1 22C3 pharmDx kit (Dako) using the Dako Autostainer Link 48 Platform (Dako). A total of 44 cases were stained with the VENTANA SP263 antibody (Ventana) test using the BenchMark ULTRA detection system (Ventana). Six cases were excluded owing to insufficient specimens (Fig. 1).

The VENTANA iScan Coreo digital pathologic slide scanner was used to scan 50 22C3-stained IHC slides and corresponding hematoxylin and eosin-stained slides to produce digital images (×400) that served as the digital material.

Establishment of Reference Values

The tumor proportion score from 0% to 100% was used to assess PD-L1–stained tissue sections by two trained senior pathologists in a double-blind independent approach to establish reference values. Any discrepant cases were assessed by two pathologists using a multiheaded microscope.

Interpreting Pathologists

A total of 20 pathologists from 20 different medical centers throughout the country were selected to represent a range of pathologists' experience, reflecting a realistic distribution of pathologists. The mean age of the interpreting pathologists (IPs) was 36 years old (range: 28–47 y), with a median of 11 years of experience (range: 5–22 y). There were one chief pathologist, six deputy chief pathologists, 12 attending pathologists, and

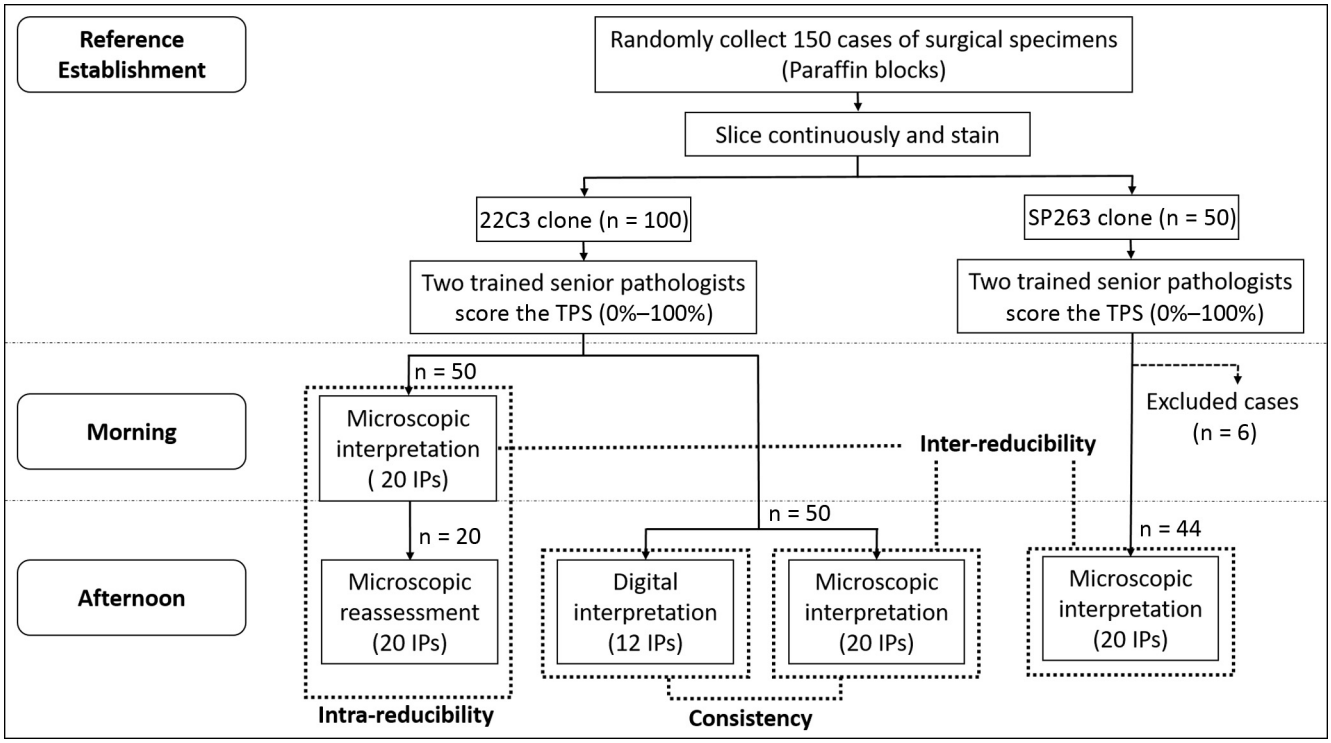

Figure 1. Flow diagram revealing the study design. IP, interpretation pathologist; TPS, tumor proportion score.

one resident pathologist, among whom 17 had received 22C3 training, of whom eight had also trained for SP263 at the same time. One pathologist had only received the SP263 training.

Scoring of PD-L1 Assays

The eligible slides were read in random order by 20 IPs, who interpreted the tumor proportion score from 0% to 100% by a double-blind, independent method. The IPs were blinded to their previous interpretations and to those of the other IPs. Staining of any intensity that was complete or partial on the tumor membrane (at a level no $<1\%$) was considered to be positive. The results of 22C3 were analyzed on the basis of two cutoffs, 1% and 50%, whereas those of SP263 were 1%, 25%, and 50%.

To reduce the intraobserver and interobserver variability caused by the heterogeneity of the interpretation time, all interpretations were completed on the same day (Fig. 1). For the morning interpretations, 20 IPs interpreted 50 cases of 22C3 using light microscopes. The afternoon interpretations were performed in three parts. In the first part, 20 IPs reassessed the 20 cases they had analyzed in the morning. In the second part, 20 IPs interpreted another 50 22C3 slides using light microscopes and assessed the digital images at the same time. Owing to some uncontrollable factors, only 12 IPs completed the digital interpretations. In the third part, 20 IPs interpreted 44 SP263 slides using light microscopes.

Statistical Analysis

Statistical analyses were undertaken using the SPSS software (version 23.0; IBM Corp., Armonk, NY). The overall percentage agreement (OPA), negative percentage agreement (NPA), positive percentage agreement (PPA), and 95% confidence interval (95% CI) were used to assess the observer reproducibility. The reliability among the pathologists for binary tumor evaluation with the specific cutoffs was assessed by Fleiss' kappa (κ), interpreted as poor to fair ($\leq$0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00).[21] The consistency between microscopic and digital interpretations was assessed by the OPA and Spearman's correlation test, in which higher consistency was defined as ρ greater than or equal to 0.80.

The percentage agreement of each pathologist compared with the recognition values (as defined subsequently) was assessed as the individual percentage agreement (IPA). The results recognized by no less than half of the 20 IPs (if it was only half, the average score was applied instead) were defined as the recognition values. For each pairwise comparison among pathologists, the results (total pairs, T) were counted as concordant pairs (CPs), including negative-negative (NN) CPs, positive-positive (PP) CPs, and discordant (D) CPs.

Table 1. Intraobserver and Interobserver Reproducibility of the 22C3 Assay

	Intraobserver (N = 400)[a]		Interobserver (N = 19,000)[b]	
Measurements	1%	50%	1%	50%
CPs	368 (92.0%)	356 (89.0%)	16,468 (86.7%)	16,948 (89.2%)
Negative-negative	35 (8.8%)	281 (70.3%)	3940 (20.7%)	12,179 (64.1%)
Positive-positive	333 (83.2%)	75 (18.7%)	12,528 (66.0%)	4769 (25.1%)
DCPs	32 (8.0%)	44 (11.0%)	2532 (13.3%)	2052 (10.8%)
Measures of agreement (95% CI)				
OPA (%)	92.0 (89.3-94.7)	89.0 (85.9-92.1)	86.7 (86.2-87.1)	89.2 (88.8-89.6)
NPA (%)	68.6 (55.9-81.4)	92.7 (89.8-95.7)	75.7 (74.5-76.8)	92.2 (91.5-93.0)
PPA (%)	95.4 (93.2-97.6)	77.3 (69.0-85.7)	90.8 (90.3-91.3)	82.3 (81.7-82.9)

[a]N = 20 (the number of IPs) × 100 (the number of cases).
[b]N = C_{20}^{2} (the number of comparison pairs of each case) × 100 (the number of cases).
CI, confidence interval; CP, concordant pair; DCP, discordant CP; IP, interpretation pathologist; NPA, negative percentage agreement; OPA, overall percentage agreement; PPA, positive percentage agreement.

Table 2. The Reliability Among the Pathologists for Binary Tumor Evaluations With the Specific Cutoff Points

PD-L1 Clone/Cutoff, %	Fleiss' κ	Interpretation
22C3		
1	0.67	Substantial
50	0.75	Substantial
SP263		
1	0.7	Substantial
25	0.46	Moderate
50	0.54	Moderate

PD-L1, programmed cell death-ligand 1.

The OPAs, NPAs, and PPAs were calculated as follows:

$$\mathrm{OPA} = (\mathrm{NN} + \mathrm{PP}) / \mathrm{T}$$

$$\mathrm{NPA} = 2 \times \mathrm{NN} / (2 \times \mathrm{NN} + \mathrm{DCP})$$

$$\mathrm{PPA} = 2 \times \mathrm{PP} / (2 \times \mathrm{PP} + \mathrm{DCP}).$$

Results

Intraobserver Reproducibility of the 22C3 Assay

For the cutoffs of 1% and 50%, there were 368 and 16,948 CPs, resulting in OPAs of 92.0% (89.3%–94.7%) and 89.0% (85.9%–92.1%), respectively (Table 1). There were four cases (1%–5%) and six cases (35%–60%) for which no less than half of the IPs had inconsistent results with the reference value in at least one assessment.

Interobserver Reproducibility of the 22C3 Assay

For the cutoff of 1%, there were 16,468 CPs, resulting in an OPA of 86.7% (86.2%–87.1%). In 50% of the cases (50 of 100), the results of the 20 IPs were completely consistent and agreed with the reference value. For the cutoff of 50%, there were 16,948 CPs, resulting in an OPA of 89.2% (88.8%–89.6%) (Table 1). In 66% of the cases (66 of 100), the results of the 20 IPs were completely consistent and agreed with the reference value. The reliability among the pathologists for binary tumor evaluations both at the cutoffs of 1% and 50% revealed substantial agreement ($\kappa = 0.67$ and $\kappa = 0.75$, respectively) (Table 2).

For the cutoff of 1%, there were 11 cases in which the recognition value was inconsistent with the reference value (range: 0%–5%). For the cutoff of 50%, there were 10 cases in which the recognition value was inconsistent with the reference value (range: 40%–60%). These special cases were all close to the specific cutoffs, 1% or 50%.

A total of 100 22C3 stained slides were divided into two for interpretation in the morning and afternoon. For the cutoff of 1%, there were 8025 and 8443 CPs, resulting in OPAs of 84.5% (83.7%–86.6%) and 88.9% (88.2%–89.5%), respectively. For the cutoff of 50%, there were 8228 and 8720 CPs, resulting in OPAs of 86.6% (85.9%–87.3%) and 91.8% (91.2%–92.3%), respectively.

Interobserver Reproducibility of the SP263 Assay

For the cutoff of 1%, there were 7708 CPs, resulting in an OPA of 92.2% (91.6%–92.8%). In 72.7% of the cases (32 of 44), the results of the 20 IPs were completely consistent and agreed with the reference value. For the cutoff of 25%, there were 6105 CPs, resulting in an OPA of 73.0% (72.1%–74.0%) (Table 3). In 29.5% of the cases (13 of 44), the results of the 20 IPs were completely consistent and agreed with the reference value. For the cutoff of 50%, there were 6849 CPs, resulting in an OPA of 81.9% (81.1%–82.8%). In 34.1% of the cases (15 of 44), the results of the 20 IPs were completely consistent and agreed with the reference value. The reliability among the pathologists for a binary

Table 3. Interobserver Reproducibility of the SP263 Assay

	SP263 (N = 8360)[a]		
Measurements	1%	25%	50%
CPs	7708 (92.2%)	6105 (73.0%)	6849 (81.9%)
Negative-negative	947 (11.3%)	3157 (37.8%)	5353 (64.0%)
Positive-positive	6761 (80.9%)	2948 (35.2%)	1496 (17.9%)
DCPs	652 (7.8%)	2255 (27.0%)	1511 (19.1%)
Measures of agreement (95% CI)			
OPA (%)	92.2 (91.6-92.8)	73.0 (72.1-74.0)	81.9 (81.1-82.8)
NPA (%)	74.4 (72.0-76.8)	73.7 (72.4-75.0)	87.6 (86.8-88.5)
PPA (%)	95.4 (94.9-95.9)	72.3 (71.0-73.7)	66.4 (64.5-68.4)

[a]N = C_{20}^2 (the number of comparison pairs of each case) × 44 (the number of cases).
CI, confidence interval; CP, concordant pair; DCP, discordant CP; NPA, negative percentage agreement; OPA, overall percentage agreement; PPA, positive percentage agreement.

Table 4. Interobserver Reproducibility of Assessment of the 22C3 Assay in Microscopic and Digital Interpretations

	22C3 (N = 3300)[a]			
	Microscopic Interpretation		Digital Interpretation	
Measurements	1%	50%	1%	50%
CPs	2957 (83.5%)	2997 (90.8%)	3050 (92.4%)	3005 (91.1%)
Negative-negative	681 (20.6%)	1823 (55.2%)	667 (20.2%)	1794 (54.4%)
Positive-positive	2276 (62.9%)	1174 (35.6%)	2383 (72.2%)	1211 (36.7%)
DCPs	343 (16.5%)	303 (9.2%)	250 (7.6%)	295 (8.9%)
Measures of agreement (95% CI)				
OPA (%)	83.5 (82.2-84.7)	90.8 (89.8-91.8)	92.4 (91.5-93.3)	91.1 (90.1-92.0)
NPA (%)	79.9 (77.2-82.6)	92.3 (91.2-93.5)	84.2 (81.7-86.8)	92.4 (91.2-93.6)
PPA (%)	93.0 (92.0-94.0)	88.6 (86.9-90.3)	95.0 (94.2-95.9)	89.1 (87.5-90.8)
Kappa	0.73 (0.70-0.75)	0.81(0.79-0.83)	0.79 (0.77-0.82)	0.82 (0.79-0.83)

[a]N = C_{12}^2 (the number of comparison pairs of each case) × 50 (the number of cases).
CI, confidence interval; CP, concordant pair; DCP, discordant CP; NPA, negative percentage agreement; OPA, overall percentage agreement; PPA, positive percentage agreement.

tumor evaluation revealed substantial agreement at the cutoff of 1% ($\kappa = 0.70$) and moderate agreement at the cutoffs of 25% and 50% ($\kappa = 0.46$ and $\kappa = 0.54$, respectively) (Table 2).

For the cutoff of 1%, there were two cases in which the recognition value was inconsistent with the reference value (range: 2%–5%). For the cutoff of 25%, there were five cases in which the recognition value was inconsistent with the reference value (range: 15%–35%). For the cutoff of 50%, there were three cases in which the recognition value was inconsistent with the reference value (range: 50%–60%). These special cases were all close to the specific cutoffs, 1%, 25%, or 50%.

The Consistency Between Microscopic and Digital Interpretations and the Interobserver Reproducibility of These Interpretations

For the cutoff of 1%, the OPAs of the microscopic and digital interpretations were 83.5% (82.2%–90.8%) and 92.4% (91.5%–93.3%), respectively. For the cutoff of 50%, the OPAs were 90.8% (89.8%–91.8%) and 91.1% (90.1%–92.0%), respectively (Table 4). The consistency between microscopic and digital interpretations (interobserver reproducibility) revealed OPAs of 93.5% (91.5%–95.5%) and 92.0% (89.8%–94.2%), respectively. Microscopic and digital interpretations were consistent ($\rho = 0.83$) at the cutoffs of 1% and 50% (Table 5). In most inconsistent cases, the reference values were close to the specific cutoffs, 1% or 50%.

The Influence of Professional Titles, Specialty, or the Number of Working Years on the Consistency of Interpretation

For the interpretations of PD-L1 stained with the 22C3 clone, the median IPA of the 20 IPs was 90.0% (range: 85%–95%) at the cutoff of 1%. There were seven IPs whose IPAs were lower than 90%, including one chief pathologist, two deputy chief pathologists, and four attending pathologists. The lowest IPA was that of an

Table 5. The Consistency Between Microscopic and Digital Interpretations

	22C3 (N = 600)[a]	
Measurements	1%	50%
CPs	561 (93.5%)	552 (92.0%)
Negative-negative	130 (21.7%)	332 (55.3%)
Positive-positive	431 (71.8%)	220 (36.7%)
DCPs	39 (6.5%)	48 (8.0%)
Measures of agreement (95% CI)		
OPA (%)	93.5 (91.5-95.5)	92.0 (89.8-94.2)
P	0.83 (0.77-0.88)	0.83 (0.77-0.88)

[a]N = 12 (the number of IPs) × 50 (the number of cases).
CI, confidence interval; CP, concordant pair; DCP, discordant CP; OPA, overall percentage agreement.

attending pathologist who was not trained in interpreting 22C3 antibody staining, whereas the other six pathologists were trained. Only two of these seven IPs had subspecialized in lung carcinoma. The median IPA was 93.0% (range: 83%–98%) at the cutoff of 50%, with three IPs having a level lower than 90%, including two attending pathologists and one resident pathologist, who had all received 22C3 training, whereas two had subspecialized in lung carcinoma.

For the interpretations of PD-L1 stained with the SP263 clone, the median IPA of the 20 IPs was 95.5% (range: 88.6%–100%) at the cutoff of 1%. There were two IPs whose IPAs were lower than 90%, including one deputy chief pathologist and one attending pathologist, both of whom had subspecialized in lung carcinoma but had not undergone training for the SP263 assay. The median IPA was 79.5% (range: 70.5%–88.6%) at the cutoff of 25%, with 11 IPs lower than 80%, including three deputy chief pathologists, seven attending pathologists, and one resident pathologist, with five IPs having received SP263 training and two of these IPs having subspecialized in lung carcinoma. The median IPA was 93.2% (range: 54.5%–100%) at the cutoff of 50%, with three IPs lower than 80%, including one deputy chief pathologist and two attending pathologists. The deputy chief pathologist had received SP263 training but was not subspecialized in lung carcinoma, and the attending pathologists were subspecialized in this but had not undergone training for SP263.

Discussion

On the basis of the results of the CheckMate, KEYNOTE, OAK, and PACIFIC trials, PD-L1 expression is currently used for immunotherapy in patients with advanced NSCLC, in whom accurate pathologic assessments are of great importance, especially for the companion diagnostic assays. Several studies have revealed that pathologic assessments of tumor cell scoring in NSCLC were highly reproducible. However, in some studies, the number of pathologists or cases enrolled was too small; thus, high reproducibility could easily be identified. Moreover, almost all the pathologists in these studies had been trained in interpreting the corresponding assays, which does not reflect actual diagnostic practice.[17-19] In contrast, the pathologists selected in this study were from different medical centers, with diversity in their training and experience.

In a study for NSCLC by Cooper et al.,[18] five pathologists assessed 60 22C3 stained samples to determine the intraobserver reproducibility, with OPAs of 89.7% and 91.3% being reported for the cutoffs of 1% and 50%, respectively. Although we also observed high intraobserver reproducibility, it was higher for the cutoff of 1%, which could be attributed to the case selection with more cases having the level of PD-L1 of approximately 50% but far away from 1%. In addition, most pathologists (17 of 20) in this study had undergone training for the 22C3 assay. Interobserver reproducibility of 22C3 was similar to that in other studies, which revealed high reproducibility among trained pathologists, but it was lower at the cutoff of 1% than in other studies.[10,11,17] Although the task of interpreting findings may be more burdensome in the afternoon, the interobserver reproducibility in the afternoon was not worse than that in the morning and was in fact actually slightly higher. Although the fatigued state may be a factor influencing the interpretation of findings, fatigue had little influence in this study, and this is the ordinary state of pathologic work in the People's Republic of China. In terms of the reliability among the pathologists for binary tumor evaluations at the cutoffs of 1% and 50%, there was substantial agreement, which was mainly accomplished because most IPs had undergone systematic training. In contrast, interobserver reproducibility was lower for SP263, with OPAs of 73.0% and 81.9% for the cutoffs of 25% and 50%, respectively. These were lower than that at the cutoff of 1%, which contrasts with the findings of earlier studies. In addition, the reliability among the pathologists for binary tumor evaluations revealed substantial agreement at the cutoff of 1% and

moderate agreement at the cutoffs of 25% and 50%. These results could have been caused by the case selection and distribution in our study; more importantly, 17 of 20 IPs had undergone 22C3 training, whereas only nine of 20 had for SP263. Therefore, the degrees of agreement at the cutoffs of 1% and 50% are higher than 25%. In addition, the interpretation of 25% is more subjective than the other two cutoffs, which is similar to the case for the interpretation of Ki-67 status, using 14% as the cutoff. Combining the results for the interpretation of PD-L1 stained with two antibodies, the cases in which the recognition values were inconsistent with the reference values reflected that the levels of both intraobserver and interobserver reproducibility were lower for the cases around the specific cutoffs despite most pathologists having undergone training for the interpretation. Although there are few cases near the threshold in practice, accurate interpretation of such cases would directly affect the therapeutic choice. Thus, the training for cases around the specific cutoffs could play a crucial role in improving the intrareproducibility or interreproducibility and providing accurate guidance for clinical treatments. Of course, in this context, there are various pitfalls and challenges, to which attention should be paid, including staining of macrophages or other immune cells, incomplete and/or weak membrane staining, and concurrent cytoplasmic staining.

With the development and increasing spread of digital pathology, greater attention has been paid to the feasibility of digital diagnostics instead of optical diagnostics; however, there are limited designs to compare the consistency between digital diagnostics and optical diagnostics. Hence, in this study, we explored the consistency between the microscopic and digital interpretations of PD-L1. These revealed strong consistency at the cutoffs of 1% and 50% ($\rho = 0.83$), in agreement with the results of Blueprint 2.[14] Furthermore, we did match the digital and glass slide scores from each pathologist to make the results more reliable. Interobserver reproducibility for the digital interpretation was higher than that for the microscopic interpretation, which could have been due to the ability of digital imaging to achieve a full preview and focused observation. These findings mean that it might be possible to improve the observer reproducibility through a digital scoring system. Limited by the deficiency of immunotherapy-related information, we can only make comparisons at the methodological level and cannot predict the correlation with clinical outcomes.

To explore the influence of training on the interobserver reproducibility of interpretations of PD-L1, we analyzed the agreement between trained and untrained pathologists. Because in our study most pathologists (17 of 20) had been trained in 22C3 interpretation whereas for SP263 the numbers of pathologists with and without training were similar (9:11), we only discussed the SP263. For the cutoffs of 25% and 50%, no substantial differences were found between the trained and untrained groups, but for the cutoff of 1%, interobserver reproducibility of the trained group was slightly higher than that of the untrained group, which could reflect the effect of training. As mentioned earlier, training, especially for the interpretation of specific cutoffs, is essential to improve reproducibility among pathologists.

To explore the influence of professional titles, specialty, and the number of working years on the consistency of interpretation, we selected 20 pathologists from different medical centers throughout the country. We found that pathologists with a lower IPA than most IPs included those with any professional title, regardless of the specialty or the number of working years. Therefore, it seems that these factors are not so important for interobserver reproducibility. On one hand, pathologists' previous experience may affect the interpretation of PD-L1, such as the interpretation of weak or incomplete membrane staining on HER2. On the other hand, it may replace part of the interpretation training of PD-L1 to some extent because all markers stained on the membrane are similar in interpretation. However, we still emphasize the importance of targeted training for specific markers as used as the companion diagnostics.

In conclusion, we found the following: (1) Intraobserver and interobserver reproducibility of PD-L1 IHC interpretations was high using the 22C3 clone but lower for the SP263 clone. Corresponding training, especially on cases around the specific cutoffs, is essential for marked improvement of the reproducibility. (2) There was strong consistency between the microscopic and digital interpretations for specific cutoffs on 22C3 clone staining. Digital interpretation could improve the reproducibility among pathologists. (3) Professional titles, specialty, and the number of working years had no impact on the consistency of interpretation.

Acknowledgments

This study was supported by the Health Science Promotion Project of Beijing (2018-TG-58) and the National Key Research and Development Program (2017YFC1311005). The authors thank all the following study participants from the 20 medical centers for their contributions to this project: Hua Du (Department of Pathology, Affiliated Hospital of Inner Mongolia Medical University, Hohhot, People's Republic of China); Qiqi Gao (Department of Pathology, The First Affiliated Hospital of Zhejiang University, Hangzhou, People's Republic of China); Lei He (Department of Pathology, National Geriatrics Center/Beijing Hospital, Beijing, People's

Republic of China); Lina Hu (Department of Pathology, Shanxi Bethune Hospital, Taiyuan, People's Republic of China); Peng Li (Department of Pathology, Affiliated Cancer Hospital of Sun Yat-sen University, Guangzhou, People's Republic of China); Jing Lian (Department of Pathology, Shanxi Cancer Hospital, Taiyuan, People's Republic of China); Chao Liu (Department of Pathology and Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, People's Republic of China); Xiaoyu Lu (Department of Pathology, Affiliated Tumor Hospital of Zhengzhou University, Zhengzhou, People's Republic of China); Jun Lu (Department of Pathology, Beijing Chaoyang Hospital, Capital Medical University, Beijing, People's Republic of China); Hongxue Meng (Department of Pathology, Affiliated Tumor Hospital of Harbin Medical University, Harbin, People's Republic of China); Shihong Shao (Department of Pathology, Affiliated Hospital of Qingdao University, Qingdao, People's Republic of China); Zhigang Song (Department of Pathology, No. 1 Medical Center, PLA General Hospital, Beijing, People's Republic of China); Kunkun Sun (Department of Pathology, Peking University People's Hospital, Beijing, People's Republic of China); Lu Wang (Department of Pathology, Xijing Hospital, Air Force Medical University, Xian, People's Republic of China); Weiya Wang (Department of Pathology, West China Hospital, Sichuan University, Chengdu, People's Republic of China); Xu Wang (Department of Pathology, Hebei Cancer Hospital, Shijiazhuang, People's Republic of China); Xiaofeng Xie (Department of Pathology, Shanghai Pulmonary Hospital affiliated to Tongji University, Shanghai, People's Republic of China); Dan Zhao (Department of Pathology, Beijing Chest Hospital, Capital Medical University, Beijing, People's Republic of China); Gang Zhao (Department of Pathology, Tianjin Cancer Hospital, Tianjin, People's Republic of China); and Yushuang Zheng (Department of Pathology, The First Affiliated Hospital of Soochow University, Soochow, People's Republic of China).

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [published correction appears in *CA Cancer J Clin*. 2020;70:313]. *CA Cancer J Clin*. 2018;68:394–424.
2. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N Engl J Med*. 2015;373:123–135.
3. 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;387:1540–1550.
4. 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;375:1823–1833.
5. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial [published correction appears in *Lancet*. 2017;389:e5]. *Lancet*. 2017;389:255–265.
6. Antonia SJ, Villegas A, Daniel D, et al. Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC. *N Engl J Med*. 2018;379:2342–2350.
7. Sul J, Blumenthal GM, Jiang X, He K, Keegan P, Pazdur R. FDA approval summary: pembrolizumab for the treatment of patients with metastatic non-small cell lung cancer whose tumors express programmed death-ligand 1. *Oncologist*. 2016;21:643–650.
8. Kazandjian D, Suzman DL, Blumenthal G, et al. FDA approval summary: nivolumab for the treatment of metastatic non-small cell lung cancer with progression on or after platinum-based chemotherapy. *Oncologist*. 2016;21:634–642.
9. Ning YM, Suzman D, Maher VE, et al. FDA approval summary: atezolizumab for the treatment of patients with progressive advanced urothelial carcinoma after platinum-containing chemotherapy. *Oncologist*. 2017;22:743–749.
10. U.S. Food & Drug Administration. FDA expands approval of Imfinzi to reduce the risk of non-small cell lung cancer progressing. https://www.fda.gov/news-events/press-announcements/fda-expands-approval-imfinzi-reduce-risk-non-small-cell-lung-cancer-progressing. Accessed February 16, 2018.
11. Ratcliffe MJ, Sharpe A, Midha A, et al. Agreement between programmed cell death ligand-1 diagnostic assays across multiple protein expression cutoffs in non-small cell lung cancer. *Clin Cancer Res*. 2017;23:3585–3591.
12. Scheel AH, Dietel M, Heukamp LC, et al. Harmonized PD-L1 immunohistochemistry for pulmonary squamous-cell and adenocarcinomas. *Mod Pathol*. 2016;29:1165–1172.
13. Adam J, Le Stang N, Rouquette I, et al. Multicenter harmonization study for PD-L1 IHC testing in non-small-cell lung cancer. *Ann Oncol*. 2018;29:953–958.
14. Hirsch FR, Mcelhinny A, Stanforth D, et al. PD-L1 immunohistochemistry assays for lung cancer: results from phase 1 of the Blueprint PD-L1 IHC assay comparison project. *J Thorac Oncol*. 2017;12:208–222.
15. Tsao MS, Kerr KM, Kockx M, et al. PD-L1 immunohistochemistry comparability study in real-life clinical samples: results of Blueprint phase 2 project. *J Thorac Oncol*. 2018;13:1302–1311.
16. 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;3:1051–1058.
17. Brunnström H, Johansson A, Westbom-Fremer S, et al. PD-L1 immunohistochemistry in clinical diagnostics of

lung cancer: inter-pathologist variability is higher than assay variability. *Mod Pathol*. 2017;30:1411–1421.
18. 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;23:4569–4577.
19. Phillips T, Simmons P, Inzunza HD, et al. Development of an automated PD-L1 immunohistochemistry (IHC) assay for non–small cell lung cancer. *Appl Immunohistochem Mol Morphol*. 2015;23:541–549.
20. Williams GH, Nicholson AG, Snead DR, et al. Interobserver reliability of programmed cell death ligand-1 scoring using the VENTANA PD-L1 (SP263) assay in NSCLC. *J Thorac Oncol*. 2020;15:550–555.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.