Editorial

The reproducibility of PD-L1 scoring in lung cancer: can the pathologists do better?

Giancarlo Troncone[1], Cesare Gridelli[2]

[1]Department of Public Health, University of Naples Federico II, Naples, Italy; [2]Division of Medical Oncology, "S.G. Moscati" Hospital, Avellino, Italy

Correspondence to: Giancarlo Troncone, MD, PhD. Department of Public Health, University of Naples Federico II, via Sergio Pansini 5, I-80131, Naples, Italy. Email: giancarlo.troncone@unina.it.

Provenance: This is an invited Editorial commissioned by Section Editor Dr. Hongbing Liu, MD, PhD (Department of Respiratory Medicine, Jinling Hospital, Nanjing University School of Medicine, Nanjing, China).

Comment on: 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.

Submitted Sep 29, 2017. Accepted for publication Oct 10, 2017.
doi: 10.21037/tlcr.2017.10.05
View this article at: http://dx.doi.org/10.21037/tlcr.2017.10.05

In the era of personalized medicine, the selection of advanced stages non-small cell lung cancer (NSCLC) patients for targeted treatments requires development, validation and continuous quality assessments of a wide array of laboratory assays, including both conventional and developing methodologies. While high throughput molecular testing approaches, to extensively assess genomic biomarkers of current and potential clinical value, are fascinating innovations in the field of modern oncology, traditional morpho-molecular methodologies such as fluorescent *in situ* hybridisation and immunohistochemical (IHC) techniques are still precious in the clinic to guide therapeutic interventions (1). This holds even more true, when considering the recent requirements to evaluate in NSCLC cells the checkpoint inhibitor programmed cell death ligand 1 (PD-L1) protein expression. Different primary antibody clones, raised against different epitopes (parts) of the PD-L1, are available (2). Each clone is linked to a specific treatment: clone 28-8 (Dako, Glostrup, Denmark) for nivolumab, 22C3 (Dako) for pembrolizumab, SP142 (Ventana, Tucson, AZ, USA) for atezolizumab and SP263 (Ventana) for durvalumab. Different clinical trial performed its own PD-L1 immunohistochemistry assay as a prepackaged kit of reagents running on company-specific staining platforms according to the manufacturers' instructions either on the Dako Link AS-48 (no longer available commercially) or on the Ventana Benchmark autostainer systems, adopting custom scoring-criteria for each assay (2).

While the role of predictive IHC can continue to flourish, the recent study by Cooper *et al.* points out the challenges faced by pathologists to consistently assess PD-L1 staining and the importance of continuous education and dedicated training (3). In fact, while a binary outcome determines, for example, the presence or absence of the ALK protein in neoplastic cells, the expression of the PD-L1 protein is a continuous variable, requiring adoption of tumour proportion scores (TPS) thresholds to deem individual tumour as "positive" or "negative" for PD-L1 expression. The clinical significance of a PD-L1 expression value around the cut off may be questionable, as patients who score just below a cut-off may not have a likelihood of response that is far different than those who scores just above the cut-off. Nevertheless, pathologists should bear in mind that scoring around any of the relevant thresholds could lead to an alternative clinical decision regarding therapy. Although, microscopy is inevitably subject to some variability, any effort to standardize both technical procedures of staining and pathologists' interpretation should be made.

The study by Cooper *et al.* focused on the interobserver reproducibility of pathologists' assessment of PD-L1 staining (3). To this end, special care was taken to avoid variability in any analytical factor, that in turn could have influenced the PD-L1 IHC test interpretation. The commercial assay Dako 22C3 pharmDx, used in this study,

is the only companion diagnostic (cdx) test approved by the US Food and Drug Administration (FDA) for treatment with pembrolizumab (4). The careful development of an initial prototype and then of an optimized clinically validated assay make this clinical trial assay a sensitive, specific, precise, and robust tool, with little interinstrument, interoperator, interday, interlot, and intrarun variations (5). In the Keynote studies, this test provided high value clinical utility to identify patients who benefit from treatment with pembrolizumab, either as first-line treatment (PD-L1 proportion score of at least 50%), or as second line in patients with any staining >1% (6). As suggested by the manufacturer, in each staining run, signal specificity is ensured by a number of controls; these comprise PD-L1 positive cell line slides, positive and negative in-house tissue samples and a negative control reagent (included in the kit) in used in place of the primary antibody on a sequential section of each patient to control for the background staining levels.

While the adoption of PD-L1 clinical trial validated assays on the company-specific staining platforms with an industry standard effectively limits analytical variability, a major challenge is the variability between pathologists' assessment of the immunohistochemistry slide. Previous studies have shown that the scoring of immune cells yields low concordance rates and that IHC is probably inadequate for assessment of immune cell expression independent of which assay is selected, with only the possible exception of the SP142 clone, specifically developed for evaluation of tumor associated immune cells (7). Conversely, several studies summarized in *Table 1*, consistently reported a relatively good agreement for PD-L1 scoring between the pathologists, independent of training and professional education (3,7-10).

As a matter of the fact, microscopic interpretation is a parameter that is inherently difficult to measure, objectively. To evaluate the ability of any given pathologist to correctly assess the immunostaining, a somehow arbitrary gold standard PD-L1 tumor proportion score is required for any given specimen. In absence of data on response to therapy there is need of a surrogate for "truth", to define as a gold standard the pathologist's score most likely to be correct (7). For example, in the study by Rimm *et al.*, the median pathologist's score was defined as the "truth" (7). Conversely, in the study by Cooper *et al.*, the gold standard was established as the consensus using a multi-headed microscope between two lead investigators, who had been specifically trained (3). Prior to study experiments, for all study samples the lead investigators assessed the percentage of PD-L1 positive tumor cells. Using these gold standard PD-L1 tumor proportion scores, a statistician performed a stratified randomization to prepare two sample sets each designed to assess reproducibility either around the 1% or the 50% cut points. Each sample set included an equal number (n=30) of PD-L1 positive and negative samples for a total of 60 samples, yielding 2,700 pairwise comparisons, for each set of cases. The study by Cooper *et al.*, featured a good representation of both specimens (n=120) and pathologists (n=10), thus that for each of the two cut points (1% and 50%), data were adequate to evaluate inter- and intraobserver reproducibility (3). Similarly, in the study by Rimm *et al.*, 13 pathologists from seven institutions performed the scoring on digital slides (performed centrally at the vendor's facilities) of 90 surgical specimens by using an internet connection (7). Brunnström *et al.* reported on 55 resected lung cancer cases, scored by seven pathologists (8), while in the study by Scheel *et al.*, 15 samples were centrally stained and scored independently by nine pathologists (10).

Another relevant methodological point is the type of tissue sample used for evaluating the reproducibility of PD-L1 tumor proportion score assessment. Cooper *et al.* employed tissue microarrays (TMA); while TMA is probably the best choice for method comparisons and investigation of interrater variation, this tissue preparations are not representative of the real clinical practice (3). Indeed, the evaluation of whole tissue section adds further complexities to the scoring, establishing which area to assess should be considered. In principle, the percentage of PD-L1 positive tumor cells should be evaluated relative to all viable tumor cells present in the specimen. Thus, the pathologists should carefully consider the overall tumor area not only in the positive areas but also in zones without any perceptible and convincing cell membrane staining. In this setting pathologists should be trained to separate tissue into areas of equal cell denominator, evaluate each area for intensity and PD-L1 positivity, add the percent positivity from each area and divide by the total number of areas.

The study experiment by Cooper *et al.* was conducted on 2 separate days. On the first day without any specific training, ten experienced pathologists scored the cases (3), evaluating as positive the cells with any perceptible membrane staining (partial or complete) perceived as being distinct from cytoplasmic staining irrespective of staining intensity. Most comparisons were concordant both for the 1% (84.2%) and for the 50% (81.9%) cut point sample sets. However, Cooper *et al.* underlined pathologists' difficulty

Table 1 Summary of literature study assessing agreement for PD-L1 scoring between the pathologists

Author (reference)	Antibody	Number of samples	Sample preparation	Number of pathologists	Cut points	Statistical test	Interobserver concordance	Gold standard
Cooper *et al.* (3)	22C3	60	TMA	10	>1% and >50%	OPA	84.2% for 1% cut point; 81.9% for 50% cut point	Lead investigators assessment
Brunnström *et al.* (8)	28-8 22C3, SP142, SP263	55	TMA	7	6-step system	Weighted kappa	0.71–0.96	Consensus
Rehman *et al.* (9)	SP142	35	Slides from 3 blocks	5	number percentages from 0–100%	ICC	94%	–
Rimm *et al.* (7)	28-8, 22C3, SP142, E1L3N	90	Slides	13	6-step system; >1% and >50%	ICC; Fleiss kappa; Kendall concordance coefficient	0.832–0.882; 0.537 for 1% cut point and 0.749 for 50% cut point; 0.612 for 1% cut point and 0.775 for 50% cut point	Median pathologist score
Scheel *et al.* (10)	28-8, 22C3, SP142, SP263	17	Slides	9	6-step system; 4-step system	Light's kappa	0.47–0.5; 0.6–0.8	–

OPA, overall percent agreement; ICC, interclass correlation coefficient; TMA, tissue microarrays.

when assessing positivity at the 50% cut point. In fact, inter-observer agreement was higher (Cohen's k coefficient 0.68) for the 1% cut point sample set, whereas it was only moderate (Cohen's k coefficient 0.58) for the 50% cut point sample set. Thus, pathologists mostly underscored the samples in the 50% cut point sample set, probably failing to recognize weak and/or incomplete membranous staining. As a matter of the fact only approximately 26% of the 50% PD-L1 positive samples were correctly scored by at least half of the ten pathologists (3). Thus, it seems that pathologists fail to score cells with weak membranous staining. Conversely, according to Cooper *et al.* the pathologists performed better when evaluating samples with a low PD-L1 tumor proportion score. However, this point is controversial as with lower cutoffs, such as 1%, there is a greater risk for inconsistent results than with a higher cutoff such as 50%. As an example, the very recent study by Brunnström *et al.* reported a higher variability for the ≥1% than for the 50% cutoff (8). In fact, different interpretation of very few or sometimes even single cells may lead to a case being classified as positive instead of negative or vice versa if using ≥1% as cutoff. In particular, it can be very difficult to clearly distinguish between "true positive" protein staining and "false-positive" artifact in specimens with lower percentages of positive cells, especially if the staining is faint. In this perspective, it should be pointed out that macrophages are a pitfall in evaluation of PD-L1. Usually, pulmonary macrophages are present in the alveolar space may contain pigmented particles in their cytoplasm and may stain stronger than the tumor cells. Thus, it is highly advisable that only pathologists experienced in thoracic pathology should evaluate PD-L1 staining in the clinical setting, reflecting their familiarity with the distinction between lung cancer and non-neoplastic cells.

In the study experiment by Cooper *et al.*, on the second day, two subgroups of five pathologists assessed all the samples for a second time either directly or after a brief training (3). The training consisted of a 1-hour presentation covering the biology of PD-L1, development of the assay, cellular expression, and strategies to optimally assess expression in NSCLCs. The training had little impact on the interobserver reproducibility with agreement rates of 82.3% (1% cut-off sample set) and of 81.7% (50% cut-off sample set), that were similar to those of the first assessments. Only, a positive training impact was observed for "easy" samples with a very high PD-L1 tumor proportion score (>80%). It is conceivable that the challenges relative to PD-L1 interpretation and

standardization represent a steep learning curve. To speed pathologists' education the importance of specific and dedicated training cannot be overemphasized. Pathologists should be also aware of the differences in interpretation among the different clinical trials assays. As an example, while cytoplasmatic staining should not be considered for 22C3, 28-8 and SP142 assays, scoring of cytoplasmatic signal by SP263 is acceptable. A 2-day educational course by Dako PD-L1, 22C3 pharnaDx, has been attended by a number of European pathologists (http://www.targos-gmbh.de/training-courses.html) who have been certified to assess PD-L1 22C3 pharmDx staining. Collaborative studies including several pathologists who followed this expert training for scoring PDL1 with investigation of interobserver and intraobserver variability of scores are much necessary. Interestingly Brunnström *et al.* showed that resident pathologists formally trained in evaluating the PD-L1 22C3 assay by Targos/Dako/NordiQC performed similarly to board-certified pathologists (8). A new generation of morpho-moleculary pathologists is highly required to face the challenges of predictive pathology.

Acknowledgements

None.

Footnote

Conflicts of Interest: The authors have no conflicts of interest to declare.

References

1. Pisapia P, Lozano MD, Vigliar E, et al. ALK and ROS1 testing on lung cancer cytologic samples: Perspectives. Cancer 2017;125:817-30.
2. Gridelli C, Ardizzoni A, Barberis M, et al. Predictive biomarkers of immunotherapy for non-small cell lung cancer: results from an Experts Panel Meeting of the Italian Association of Thoracic Oncology. Transl Lung Cancer Res 2017;6:373-86.
3. 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.
4. Jørgensen JT. Companion diagnostic assays for PD-1/PD-L1 checkpoint inhibitors in NSCLC. Expert Rev Mol Diagn 2016;16:131-3.
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. Pai-Scherf L, Blumenthal GM, Li H, et al. FDA Approval Summary: Pembrolizumab for Treatment of Metastatic Non-Small Cell Lung Cancer: First-Line Therapy and Beyond. Oncologist 2017;22:1392-9.
7. 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.
8. Brunnström H, Johansson A, Westbom-Fremer S, et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher thanassay variability. Mod Pathol 2017;30:1411-21.
9. Rehman JA, Han G, Carvajal-Hausdorf DE, et al. Quantitative and pathologist-read comparison of the heterogeneity of programmed death-ligand 1 (PD-L1) expression in non-small cell lung cancer. Mod Pathol 2017;30:340-9.
10. Scheel AH, Dietel M, Heukamp LC, et al. Harmonized PD-L1 immunohistochemistry for pulmonary squamous-cell and adenocarcinomas. Mod Pathol 2016;29:1165-72.

Cite this article as: Troncone G, Gridelli C. The reproducibility of PD-L1 scoring in lung cancer: can the pathologists do better? Transl Lung Cancer Res 2017;6(Suppl 1):S74-S77. doi: 10.21037/tlcr.2017.10.05

