Multicentre study on the consistency of PD-L1 immunohistochemistry as predictive test for immunotherapy in non-small cell lung cancer

Rogier Butter ,[1] Nils A 't Hart,[2] Gerrit K J Hooijer,[1] Kim Monkhorst,[3] Ernst-Jan Speel,[4] Paul Theunissen,[5] Erik Thunnissen,[6] Jan H Von der Thüsen,[7] Wim Timens,[2] Marc J van de Vijver[1,8]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jclinpath-2019-205993).

For numbered affiliations see end of article.

Correspondence to
Rogier Butter, Department of Pathology, Cancer Center Amsterdam, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, 1105 AZ, The Netherlands; r.butter@amc.uva.nl

Received 24 May 2019
Revised 23 October 2019
Accepted 16 November 2019
Published Online First 10 December 2019

ABSTRACT

Aims Investigate the impact of interlaboratory- and interobserver variability of immunohistochemistry on the assessment of programmed death ligand 1 (PD-L1) in non-small cell lung cancer (NSCLC).

Methods Two tissue microarrays (TMAs) were constructed from 50 (TMA-A) and 51 (TMA-B) resected NSCLC cases, and distributed among eight centres. Immunostaining for PD-L1 was performed using Agilent's 22C3 pharmDx Assay (pharmDx) and/or a 22C3 laboratory developed test (LDT). The interlaboratory variability of staining- and interobserver variability of scoring for PD-L1 were assessed in selected critical samples (samples at the cut-off of positivity) and non-critical samples. Also, PD-L1 epitope deterioration in time in stored unstained slides was analysed. Krippendorff's alpha values (0=maximal, 1=no variability) were calculated as measure for variability.

Results For interlaboratory variability of immunostaining, the percentage of PD-L1 positive cases among centres ranged 40%–51% (1% cut-off) and 23%–30% (50% cut-off). Alpha values at 1% cut-off were 0.88 (pharmDx) and 0.87 (LDT) and at 50% cut-off 0.82 (pharmDx) and 0.95 (LDT). Interobserver variability of scoring resulted in PD-L1 positive cases ranging 29%–55% (1% cut-off) and 14%–30% (50% cut-off) among pathologists. Alpha values were at 1% cut-off 0.83 (TMA-A) and 0.66 (TMA-B), and at 50% cut-off 0.77 (TMA-A) and 0.78 (TMA-B). Interlaboratory variability of staining was higher ($p<0.001$) in critical samples than in non-critical samples at 50% cut-off. Furthermore, PD-L1 epitope deterioration in unstained slides was observed after 12 weeks.

Conclusions The results provide insight in factors contributing to variability of immunohistochemical assessment of PD-L1, and contribute to more reliable predictive testing for PD-L1.

To cite: Butter R, 't Hart NA, Hooijer GKJ, *et al*. *J Clin Pathol* 2020;**73**:423–430.

INTRODUCTION

Immune checkpoint inhibiting therapies, which target the interaction between programmed cell death receptor-1 (PD-1) and its ligand (PD-L1) have improved the survival rates of advanced non-small cell lung cancer (NSCLC).[1–4] The mechanism of action of these therapies is to target the PD-1/PD-L1 co-inhibitory signal, which suppresses the immune response against the cancer cells after antigen recognition by T-cells.[5] Inhibition of this PD-1/PD-L1 co-inhibitory signal can result in an immune response against tumour cells.[6]

PD-1 inhibitors nivolumab and pembrolizumab, and PD-L1 inhibitor atezolizumab have been proven effective in the treatment of advanced lung cancer in several randomised clinical trials.[1–4 7 8] Recently pembrolizumab has shown to be effective in combination with chemotherapy compared with chemotherapy alone in the KEYNOTE-189 trial.[9] Also, PD-L1 inhibitors avelumab and durvalumab showed promising results in phase I trials.[10 11] Durvalumab increased progression free survival as consolidation therapy after chemotherapy in a placebo controlled trial in patients with stage III NSCLC.[12]

Predictive factors, which have been associated with response to PD-1/PD-L1 inhibiting immunotherapy, are either tumour or immunogenic related. Tumour related factors include PD-L1 expression on the tumour cells, mismatch repair deficiency and mutational load. Immunogenic related factors include inflammation associated genes, blood neutrophil and lymphocyte counts, the presence of tumour infiltrating lymphocytes and human leucocyte class I diversity.[13 14] For nivolumab, higher percentages of PD-L1 positive tumour cells are associated with increased clinical response,[1 2] while pembrolizumab can be applied in patients with advanced NSCLC that show at least 50% (first line therapy) or 1% (second line therapy) PD-L1 positive tumour cells.[3 4 15]

The expression of PD-L1 on tumour cells is assessed by immunohistochemistry on formalin fixed paraffin embedded (FFPE) tumour tissue, obtained by biopsy or in resected tumours.[16] PD-L1 scores assessed by immunohistochemical assays however may vary due to (I) intratumor heterogeneity of PD-L1 expression, (II) differences in primary antibodies, (III) signal enhancement, (IV) staining platforms (V) interobserver variability and (VI) pre-analytical variation.[17–20] Currently used anti-PD-L1 monoclonal antibodies (mAbs) include among others 22C3 (Agilent), 28–8 (Agilent), SP263 (Ventana) and SP142 (Ventana). Agilent's 22C3 is available as companion diagnostic for pembrolizumab in NSCLC, but also Ventana's SP263 is approved for use.[21 22]

Given the inconsistencies in immunohistochemical scoring for PD-L1 and its importance as a predictive test for immunotherapy in NSCLC, the aim of this study was to assess to which extent

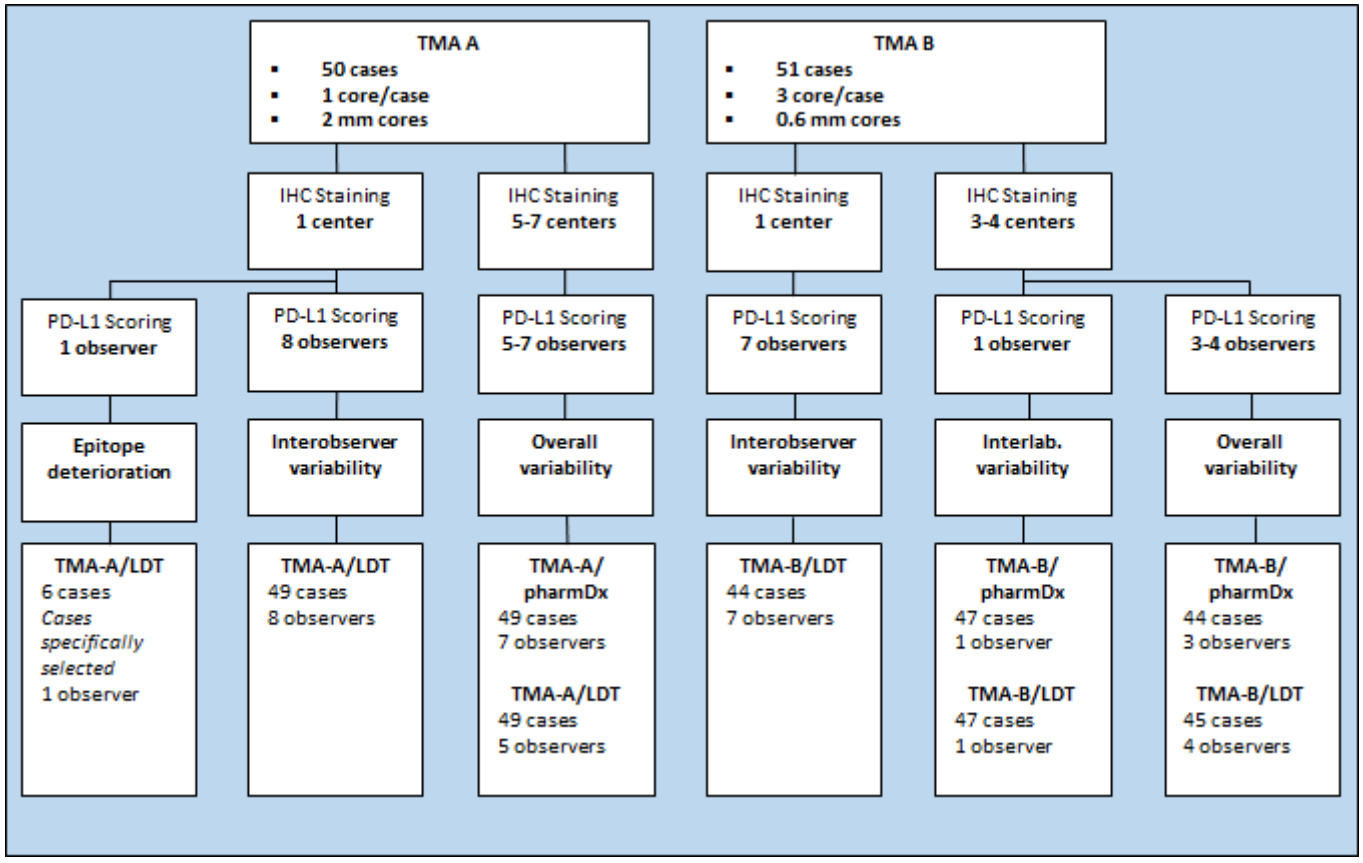

Figure 1 Study flow diagram. IHC, immunohistochemistry.

interlaboratory variability of immunohistochemical staining and interobserver variability of scoring contribute to discrepancies in PD-L1 positivity among different centres. Interlaboratory variability was also assessed in selected critical samples.[23] Furthermore, we aimed to assess PD-L1 deterioration in stored unstained slides at different time points after cutting slides from FFPE tissue blocks.

This study provides insight in factors contributing to discrepancies in immunohistochemical assessment of PD-L1 positivity, and contributes to a more robust diagnostic trajectory for the assessment of PD-L1 status of the tumour as a guiding factor for the application of immunotherapy in NSCLC.

METHODS

Study design

Ten pathologists specialised in pulmonary oncology, affiliated to eight medical centres in the Netherlands participated. Two tissue microarrays (TMAs) were constructed, consisting of NSCLC cases. These two TMAs were used for assessing variability in PD-L1 scoring among centres when immunostaining and scoring was performed in the same centre, variability of PD-L1 immunostaining among laboratories and variability of PD-L1 scoring among pathologists (figure 1).

Tissue microarrays

Two TMAs were created from FFPE tissue blocks containing pathological confirmed NSCLC tissue from resected lung specimens. Material was collected in the Amsterdam UMC, University of Amsterdam (study centre, TMA-A) and University Medical Centre Groningen, University of Groningen (TMA-B). TMA-A consisted of 50 cores (50 cases, one core of 2 mm per case). TMA-B consisted of 153 cores (51 cases, three cores of 0.6 mm per case). All selected cases were pathologically confirmed NSCLCs.

NSCLCs are PD-L1 positive (TPS ≥1%) in approximately 60% and in (TPS ≥50%) approximately 20%. We tried to approach a similar distribution in both TMAs. For that purpose, whole tissue slides from a large set of cases underwent immunostaining for PD-L1, and for each TMA cases were selected based on the PD-L1 scores obtained using whole tissue slides.

Immunohistochemistry

TMA slides were freshly sectioned at five µm and were sent to each of the participating centres. Each centre performed immunohistochemical staining for PD-L1 within 4 weeks, using antibody clone 22C3 (Agilent). Staining was performed using the standard Agilent 22C3 pharmDx Assay or a laboratory developed test (LDT) on Ventana's BenchMark ULTRA (BMU), developed by the University Medical Centre in Groningen, validated against the Agilent 22C3 pharmDx Assay. The pharmDx Assay was available in 7/8 centres and the LDT in 5/8 centres. As a result of the two TMAs and two protocols for immunohistochemical staining, four different subgroups were analysed: TMA-A/22C3 pharmDx, TMA-A/22C3 LDT, TMA-B/22C3 pharmDx and TMA-B/22C3 LDT (online supplementary figure S1).

Scoring protocol

The tumour proportion score (TPS) was defined as the proportion of tumour cells, with membranous expression of PD-L1 using mAb 22C3, according to the 22C3 pharmDx Assay instructions for use. Scores were subdivided in three categories: <1%, 1%–49% or ≥50%.

Variability among centers of PD-L1 TPS

Immunohistochemical staining and scoring was performed within the same centre, using unstained slides from both TMAs. Seven pathologists scored TMA-A/22C3 pharmDx, five TMA-A/22C3 LDT, three TMA-B/22C3 pharmDx and four TMA-B/22C3 LDT.

Interlaboratory variability of PD-L1 TPS

Slides of TMA-B were stained in each of the centres, and were sent to the study centre and PD-L1 staining was scored by one trained pathologist who was blinded for the staining protocol. Four laboratories performed immunohistochemical staining using Agilent's 22C3 pharmDx assay (TMA-B/22C3 pharmDx) and four other centres used the 22C3 LDT (TMA-B/22C3 LDT).

Interlaboratory variability of PD-L1 TPS in critical samples

The PD-L1 epitope concentration of a critical sample is around the cut-off value of a clinically validated PD-L1 immunohistochemical test.[16 23]

Selection of critical samples was performed on serial sections from TMA-B. Tissue slides underwent immunostaining with antibody 22C3 in two different dilutions: 1:25 and 1:100. Samples with major change in intensity of PD-L1 expression between these two conditions were marked as critical samples. A change in antibody dilution changes the level of PD-L1 expression in these critical cores, while PD-L1 expression in the remaining cores remained equal. Therefore, critical cores are suitable to detect differences between tests and laboratories testing for PD-L1.

Interlaboratory variability was assessed between the samples designated as critical and the remaining ("non-critical") samples.

Interobserver variability of PD-L1 TPS

One slide from TMA-A and B each was stained at the study centre using the 22C3 LDT on Ventana's BenchMark ULTRA. The stained slide was digitalized using a Philips IntelliSite Pathology Solution Ultra Fast Scanner 1.6 (Philips Digital Pathology Solutions). The digital slides were scored using a computer image. Eight pathologists scored TMA-A/22C3 LDT and seven pathologists TMA-B/22C3 LDT.

Epitope stability

The deterioration of PD-L1 epitopes on tumour cells was assessed by creating a time series of tissue slides from TMA-A. Seven PD-L1 positive slides were stained with mAb 22C3 on Ventana Benchmark Ultra after 1 day, 1 week, 4 weeks, 12 weeks and 24 weeks after tissue sections were cut. One trained, blinded pathologist scored the percentage of PD-L1 (categories: <1%, 1%–49% and ≥50%) and staining intensity (0, 1+, 2+, 3+).

Statistical analysis

IBM SPSS Statistics 24 was used to perform statistical analyses. Krippendorff's alpha was used as statistical test to express reliability rates, results ranging from 0 (no concordance) to 1 (perfect concordance). Krippendorff's alpha was selected because of its ability to correct for missing data in a dataset with multiple raters.[24 25] Also Fleiss' Generalised kappa values were calculated, which are unable to correct for missing data, but are more commonly used in literature. Chi-squared tests were performed for statistical analysis of critical samples.

RESULTS

Cases

Forty-nine cases were evaluable in TMA-A (table 1). One case was excluded because the absence of tumour tissue. Forty-seven cases were evaluable in TMA-B . Four cases were excluded from analysis because the absence of tumour tissue.

Table 1 Histology of selected cases

	TMA-A (=49)	TMA-B (=45)
Adenocarcinoma	24	19
Squamous cell carcinoma	18	20
Large cell neuroendocrine carcinoma	1	1
Large cell carcinoma	4	1
Ccarcinoid tumour	2	0
Pleomorphic carcinoma	0	4

Staining and scoring in each center

TMA-A underwent immunohistochemical staining according to the 22C3 pharmDx Assay in six centres, and was scored by seven pathologists (TMA-A/22C3 pharmDx). Four centres used the LDT on TMA-A, and was scored by five pathologists (TMA-A/22C3 LDT). One centre had two scoring pathologists. Three centres used the 22C3 pharmDx Assay on TMA-B (TMA-B/22C3 pharmDx) and four the LDT on TMA-B (TMA-B/22C3 LDT), scoring was performed by respectively three and four pathologists.

The percentage of PD-L1 positive cases was different among centres, at both the 1% and 50% cutoff value (figure 2A–D). Krippendorff's alpha values ranged between 0.83–0.85 (TMA-A/22C3 pharmDx), 0.43–0.52 (TMA-A/22C3 LDT), 0.90–0.94 (TMA-B/22C3 pharmDx) and 0.57–0.71 (TMA-B/22C3 LDT), depending on the categories in which was scored (figure 2E). Fleiss' Generalised kappa values ranged 0.73–0.83 (TMA-A/22C3 pharmDx), 0.31–0.40 (TMA-A/22C3 LDT), 0.87–0.89 (TMA-B/22C3 pharmDx) and 0.55–0.76 (TMA-B/22C3 LDT) (online supplementary table S1) in the evaluable cases, also depending on the categories.

Interlaboratory variability of PD-L1 immunostaining

Slides from TMA-B were distributed, on which four centres applied the 22C3 pharmDx Assay and four centres applied the 22C3 LDT for immunohistochemical staining. For both TMA-B/22C3 pharmDx and TMA-B/22C3 LDT 47 cases were evaluable, four cases were recorded as missing due to insufficient quality of the cores. PD-L1 positivity was different among the laboratories, depending on the cut-off values (figure 3A and B).

Krippendorff's alpha values ranged between 0.82 and 0.88 (TMA-B/22C3 pharmDx) and 0.87 and 0.95 (TMA-B/22C3 LDT), depending on the cut-off values (figure 3E). Kappa values ranged 0.63–0.85 (TMA-B/22C3 pharmDx) and 0.75–0.88 (TMA-B/22C3 LDT), depending on the cut-off values (online supplementary table S1). An example of differences in TPS of the same core, despite using the same protocol (22C3 pharmDx Assay) is shown in figure 4.

Interlaboratory variability of PD-L1 immunostaining in critical samples

Critical samples were assessed in TMA-B. Six out of 47 evaluable samples were identified as critical samples, 3 at the 1% and 50% cut-off value each. Eight centres performed immunohistochemical staining, four using the 22C3 pharmDx Assay and four the 22C3 LDT. As a result, 24 scores were granted to critical samples at each cut-off (3 samples×8 observers) and 327 scores to non-critical samples (40 samples×8 observers+1 sample×7

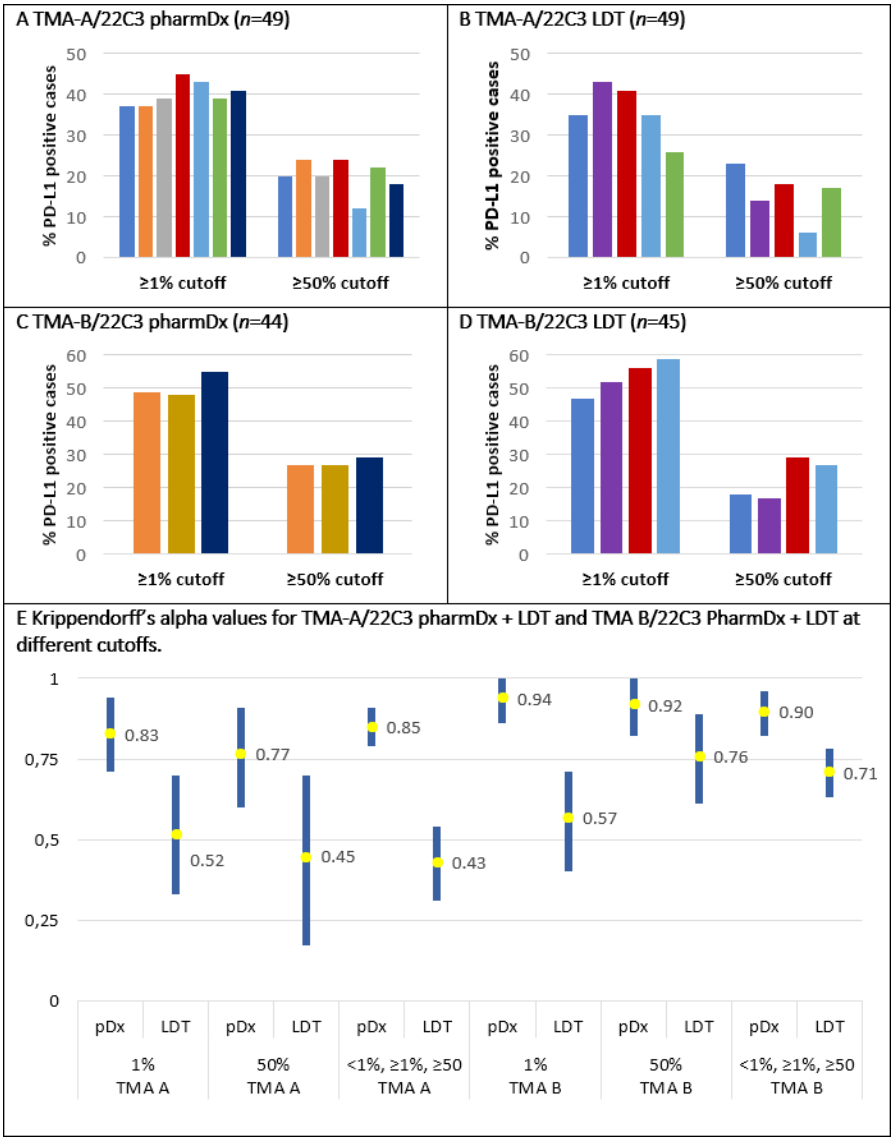

Figure 2 Immunohistochemical immunostaining and PD-L1 scoring in each centre individually. A–D show PD-L1 positive cases at different cutoffs (1% and 50%). The coloured bars represent the different centres: Blue=centre A, purple=centre B, orange=centre C, grey=centre D, red=centre E, light blue=centre F(1), green=centre F(2), brown=centre G and dark blue=centre hour. not all centres participated in evaluating both TMA-A and B, therefore the number of centre differs between TMA-A and B, centre F had two participating pathologists. E shows Krippendorff's alpha values for TMA-A and B, for both 22C3 pharmDx and 22C3 LDT, at cut-off values 1% and 50%, and when three categories were used:<1%, 1%–49% (in graph:≥1%) and ≥50%.

observers). At the 1% cut-off value, TPSs were concordant in 22/24 (92%) critical samples and in 314/327 (96%) non-critical samples (χ^2, p=0.3). At the 50% cut-off value, 19/24 (79%) critical samples were concordant and 322/327 (98%) non-critical samples were concordant (p<0.001) (online supplementarty table S2).

Interobserver variability of PD-L1 scoring

Eight pathologists scored TMA-A/22C3 LDT of which 49 cases were evaluable. Seven pathologists scored TMA-B/22C3 LDT, of which 44 cases were evaluable (figure 3C–3D).

Alpha values ranged 0.77–0.86 (TMA-A/22C3 LDT) and 0.66–0.78 (TMA-B/22C3 LDT), depending on the categories (figure 3E). Kappa values ranged 0.77–0.84 (TMA-A/22C3 LDT) and 0.62–0.76 (TMA-B/22C3 LDT) (online supplementary table S1).

Two cases (three cores each), which resulted in discordant scores among the pathologists, are shown in figure 5. Background staining and macrophage enhancement may have contributed to the discordances of the scores.

Epitope stability

Seven cores were evaluable for epitope deterioration in stored unstained slides (figure 6). All cores were PD-L1 positive at day 1. Most cores showed a decreased TPS and/or staining intensity at or after 12 weeks (table 2).

DISCUSSION

This study demonstrates that both interlaboratory variability of PD-L1 immunostaining and interobserver variability of PD-L1 scoring both contribute to different results for final PD-L1 positivity among centres. In particular, interobserver comparisons

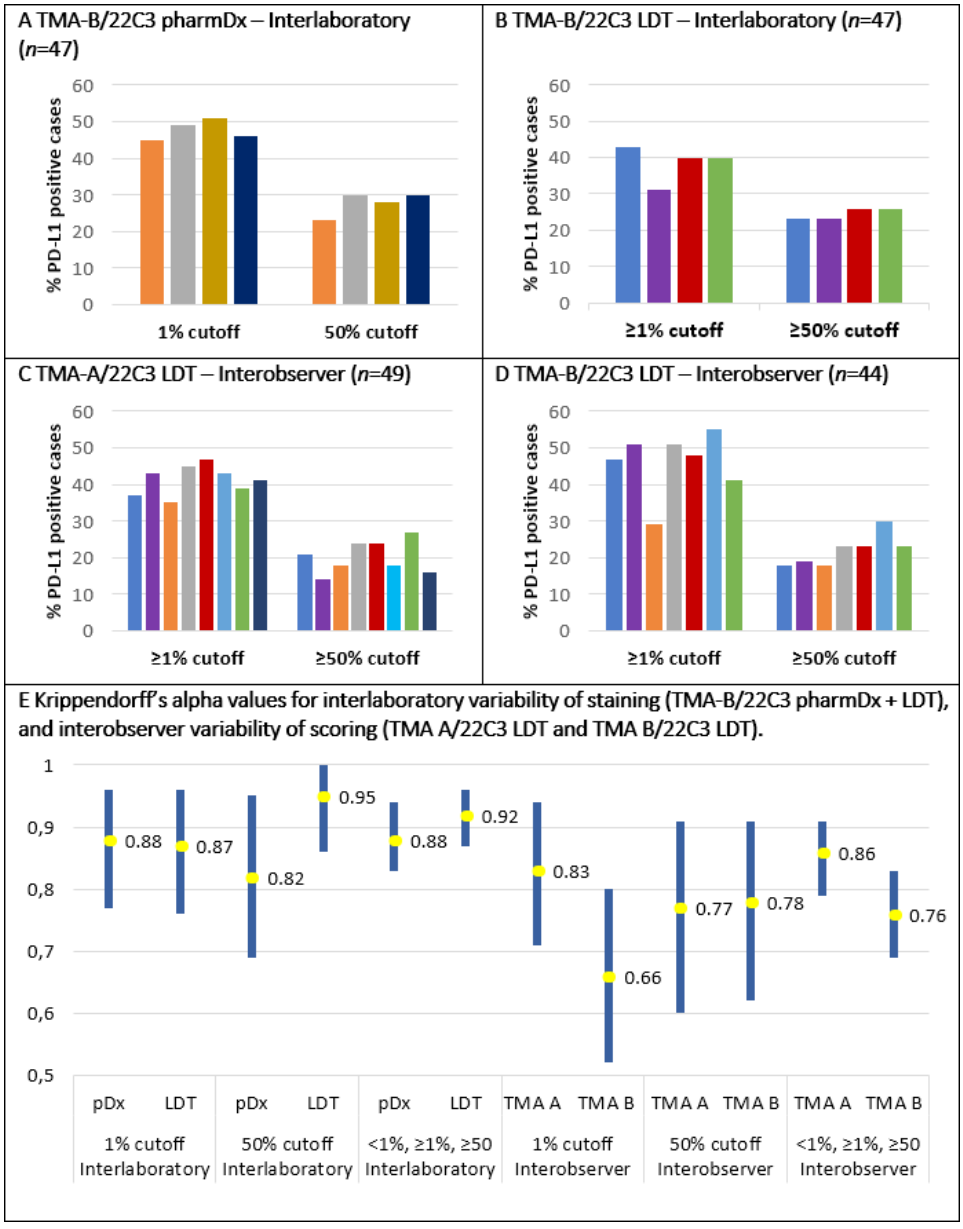

Figure 3 Interlaboratory variability of immunostaining and interobserver variability of PD-L1 scoring. (A,B) Variability of the percentage of PD-L1 positive cases of interlaboratory variability. (C,D) The variability of the percentage of PD-L1 positive cases of interobserver variability. The coloured bars represent the different centres: blue=centre A, purple=centre B, orange=centre C, grey=centre D, red=centre E, light blue=centre F(1), green=centre F(2), brown=centre G and dark blue=centre hour. (E) Krippendorff's alpha values of interlaboratory variability of staining and interobserver variability of scoring at cut-off values <1%, 1%–49% (in graph: ≥1%) and ≥50%. LDT,laboratory developed test; PD-L1, programmed death ligand 1; TMA,tissue microarray.

showed the largest variation, both in the 22C3 pharmDx Assay an the 22C3 LDT. Also, test results when both the immunohistochemical staining and PD-L1 scoring was performed within the same centre, showed more variability among centres which use the LDT than centres which use the 22C3 pharmDx Assay.

Three previous studies, dedicated to antibody development for use in NSCLC (22C3, SP142 and SP263), performed analyses in which PD-L1 scoring results were compared when immunohistochemical staining and PD-L1 scoring was performed within the same centre.[15 22 26] Roach *et al* performed immunohistochemical staining with the 22C3 pharmDx Assay on 36 NSCLC FFPE tissue specimens in three different centres. Subsequent scoring was performed by one pathologist per centre.[15] An overall percent agreement of 88.3% was found at a cut-off value of 50% PD-L1 positive tumour cells, which is consistent with our results. Two studies investigated Ventana's SP142 (Vennapusa *et al*) and SP263 assays (Rebelatto *et al*): 28 FFPE NSCLC tissue slides (biopsy/resections specimens) and 14 FFPE NSCLC whole tissue samples were analysed, respectively. For both studies, three separate centres performed immunohistochemical staining in which two independent pathologists per centre scored for PD-L1. Vennapusa *et al* applied a combined PD-L1 TPS and tumour infiltrating immune cells staining (TC/IC, score 1–3). Concordance scores were 93.5% (TC1/IC1), 91.2% (TC2/IC2) and 93.2% (TC3/IC3) between the laboratories.[26] Rebelatto *et al* found concordance rates of 93.3% in positive samples and 79.5% in negative samples, using 25% PD-L1 tumour cell staining as cut-off value.[22] The results of the abovementioned

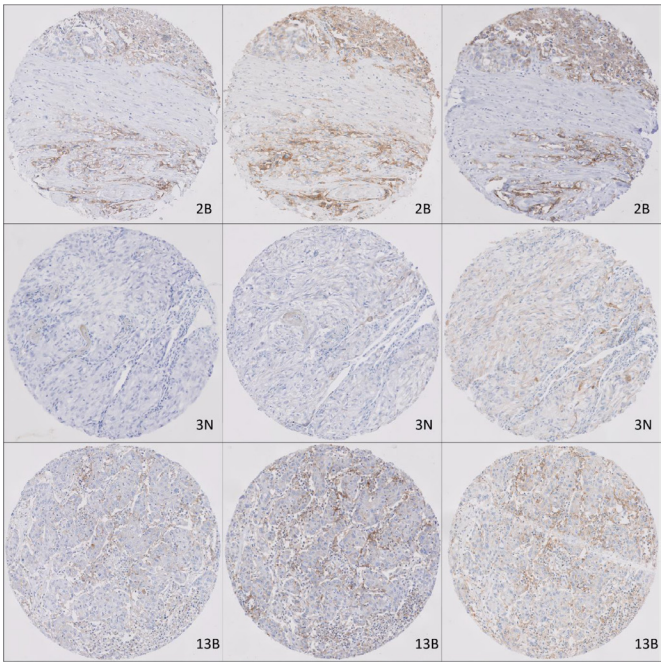

Figure 4 Interlaboratory variation of immunohistochemical staining for PD-L1 of three cores from TMA-B/22C3 pharmDx. of each case (2B, 3N and 13B), three cores which underwent immunostaining in different laboratories are shown.

studies and our results indicate that clinical decision-making for immunotherapy is affected by the methodology for PD-L1 staining and scoring. Differences may be attributable to inter-laboratory and interobserver variability. In our study, we found that the 22C3 pharmDx Assay resulted in less variability than the uniformly applied LDT.

Interlaboratory variability of staining revealed alpha values of 0.88 (1% cut-off), 0.82 (50% cut-off) and 0.88 (three categories) for laboratories which applied the 22C3 pharmDx Assay. For laboratories which applied the 22C3 LDT, alpha values of 0.87 (1% cut-off), 0.95 (50% cut-off) and 0.92 (three categories) were found. The results found for the 22C3 pharmDx Assay are consistent with a study by Scheel *et al*, which found kappa values of 0.87 (1% cut-off), 0.89 (50% cut-off) and 0.83

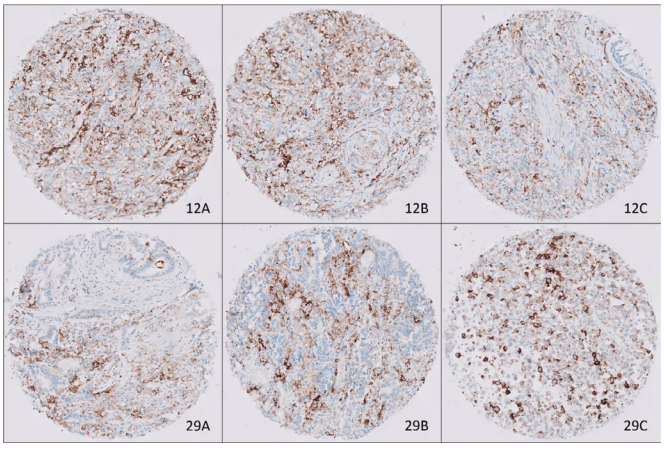

Figure 5 Interobserver variation of PD-L1 scoring. two cases (12 and 29) are shown, three cores each. these cases resulted in very different PD-L1 scores among seven pathologists: 3x<1%, 3×1%–49% and 1x>50% (case 12), and 4x<1% and 3×1%–49% (case 29).

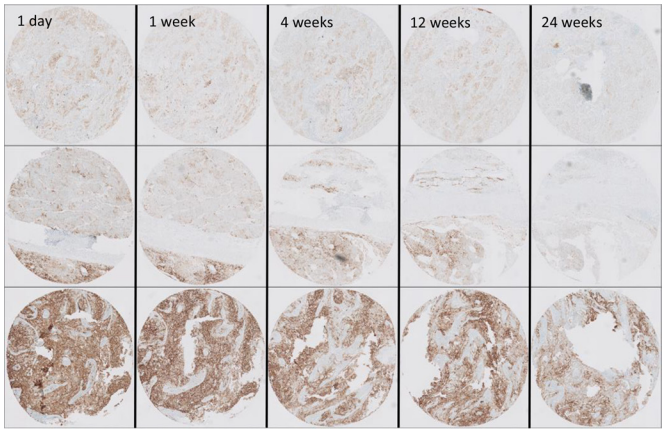

Figure 6 Pd-L1 epitope deterioration shown in three cases from TMA-A/22C3 LDT at time points at 1 day, 1 week, 4 weeks, 12 weeks and 24 weeks each row represents one case.

(three categories) in samples stained with the 22C3 pharmDx Assay.[20] These results indicate that even among laboratories which use the same staining technique and protocol results can be different. Also, a validated and uniformly applied LDT, as in our study, can yield reproducible results among laboratories at least as good as the 22C3 pharmDx Assay. It should, however, be noted that a limitation of this analysis is that scoring for inter-laboratory variability was performed by a single pathologist. Therefore, intraobserver variability can at least partly reflect the variability in our study.[27 28] Also, part of the variability among centres might be caused by PD-L1 heterogeneity among the different TMA slides, as a result of cutting slides from the TMA at different depths.

The evaluation of critical samples may be of added value.[16 23 29] Our results show that in the critical samples we selected, at the 50% cut-off value significant more variability was seen in critical samples than in non-critical samples. However, at the 1% cut-off value, no significant difference was observed, indicating that IHC assays underscored more than on the critical samples alone.

On interobserver variability of scoring, eleven studies provide evidence, and eight studies did so in the context of mAb 22C3.[15 19 22 26 28 30–35] Our study forms an addition to these previous studies because of the comparison of 22C3 as a commercial clinically validated assay to an uniformly applied LDT. Furthermore, the samples used to investigate interobserver variability have also been used for evaluation of interlaboratory variability of staining and when staining and scoring was performed in within the same centre.

Roach *et al* found an overall concordance rate of 96.4% among three observers at a cut-off of 50% PD-L1 positive tumour cells and Rimm *et al* found an interclass correlation coefficient of 0.88 among 16 pathologists and 90 samples (both using 22C3 pharmDx Assay).[15 30] Munari *et al* found a Cohen's kappa value of 0.77 between two pathologists who used the 22C3 pharmDx Assay in 198 NSCLC cases.[35] Another study by Scheel *et al* found Light's kappa's of 0.74 and 0.66 at the 1% and 50% cut-off values, scored by nine pathologists, using whole slides of lung squamous cell carcinoma and adenocarcinoma stained with the 22C3 pharmDx Assay.[19] Two sample sets of 60 NSCLC cases (Cooper *et al*), stained using the 22C3 pharmDx Assay and scored by 10 pathologists led to an overall percent agreement of 84.2% and a Cohen's kappa of 0.68 at the 1% cut-off and 81.9% and a kappa of 0.58 at the 50% cut-off value.[28] The kappa values between Scheel *et al* and Cooper *et al* overlap, however, the overall percent agreement in

Table 2 Epitope deterioration at 1 day (d), 1, 4, 12 and 24 weeks (w)

		Tumour proportion score (intensity) at certain time point							
Core	1 d	1 w		4 w		12 w		24 w	
1	1%–49% (2+)	1%–49% (1+)	↓	1%–49%(1+)	~	1%–49% (1+)	~	<1% (1+)	↓
2	1%–49% (2+)	1%–49% (1+)	↓	1%–49% (2+)	↑	1%–49% (1+)	↓	1%–49% (1+)	~
3	1%–49% (1+)	≥50% (1+)	~	≥50% (1+)	~	≥50% (1+)	~	≥50% (1+)	~
4	≥50% (3+)	≥50% (3+)	~	≥50% (3+)	~	≥50% (3+)	~	≥50% (2+)	↓
5 (1/2)	≥50% (1+)	≥50% (1+)	~	≥50% (1+)	~	≥50% (1+)	~	≥50% (1+)	~
5 (2/2)	1%–49% (3+)	1%–49% (3+)	~	1%–49% (3+)	~	1%–49% (2+)	↓	1%–49% (2+)	~
6	≥50% (2+)	≥50% (2+)	~	≥50% (2+)	~	≥50% (2+)	~	≥50% (1+)	↓

An unchanged PD-L1 score or staining intensity is indicated by ~, a decrease by ↓ and increase by ↑. Three categories were used for PD-L1 expression:<1%, 1%–49% and ≥50%. Three categories were used for intensity: 1+ to 3+; 3+ being most intense.
PD-L1, programmed death ligand 1.

Scheel *et al* is much lower due to the application of six categories in the scoring method, instead of using one cut-off value. Weighted kappa values ranging from 0.71 to 0.95 were found among seven observers who scored 55 NSCLC samples stained with 22C3 in a study by Brunström *et al*. However, only kappa values between pairs of pathologists were calculated and no comparison of all scored samples was made.[31]

These studies and our results indicate a wide variety of PD-L1 scoring results between pathologist, independent of the staining protocol. Variability rates tend to improve when one cut-off value is used, instead of a multistep scoring system. Cut-off values based on the intended use in the respective instructions should be used for clinical decision-making.

TMA-B had higher alpha values compared with TMA-A when immunostaining and scoring was performed in one centre. Especially for 22C3 LDT, four labs scored for both TMAs, resulting in higher alpha values for TMA-B. However, when only scoring (and not also staining) was assessed, TMA-A had higher alpha values for interobserver variability compared with TMA-B. In view of this, results were not TMA dependent.

The final objective of this study was to assess epitope deterioration in time of unstained FFPE slides. For research purposes or quality assurance programmes unstained slides are distributed among centres, therefore epitope stability after slide preparation is important. In our included samples, PD-L1 epitopes showed deterioration after 12 weeks. The current instructions for use of 22C3 pharmDx assay states an maximum interval of 6 months when slides are stored at 4°C to 25°C. The current study has been performed with a 22C3 LDT assay, which may account for the difference. Rebelatto *et al* reported no epitope deterioration after ten months in NSCLC samples, using mAb SP263.[22] To our knowledge, no further evidence of epitope stability detected with mAb 22C3 has been reported.

The results of this study and results from previous studies indicate that the most limiting factor in PD-L1 tumour proportion scoring is interobserver variation. Improvement may be achieved by training of pathologists, however, Cooper *et al* showed that pathologist training had no or little effect on the improvement of interobserver variability.[28]

This study provides results on the application of 22C3 as commercial assay or as LDT. These results cannot be extrapolated to other antibodies, which assess PD-L1 expression in NSCLC, since numerous studies have shown that differences in scores exist between antibodies in various combinations.[19 30 31 36–42]

In conclusion, both interlaboratory variability of staining and interobserver variability of scoring contribute to discordances in PD-L1 positivity among centres, especially in cases where the PD-L1 expression is around the cut-off. New biomarkers can contribute to a more reliable patient selection for the application of immunotherapy in NSCLC.

Take home messages

- Protocolised predictive immunohistochemical scoring for programmed death ligand 1 (PD-L1) can still lead to different decisions about the application of immunotherapy.
- Both interlabortatory and interobserver variability contribute to different PD-L1 test outcomes.
- Difficulties are primarily seen around the 1% and 50% cut-off values.
- PD-L1 epitopes deteriorate after 12-week storage as unstained slide.

Author affiliations
[1]Department of Pathology, Cancer Center Amsterdam, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands
[2]Department of Pathology and Medical Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[3]Department of Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands
[4]Department of Pathology, GROW-School for Oncology and Developmental Biology, Maastricht University Medical Center, Maastricht, The Netherlands
[5]Department of Pathology, Zuyderland Medical Center, Heerlen, The Netherlands
[6]Department of Pathology, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[7]Department of Pathology, Erasmus University Medical Center, Erasmus University Rotterdam, Rotterdam, The Netherlands
[8]Department of Pathology, Cancer Center Amsterdam, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Handling editor Cheok Soon Lee.

Acknowledgements We would like to thank Vincent Smit, MD PhD and Tjalling Bosse, MD PhD (pathologists at Leiden University Medical Center) for their contribution to this study.

Contributors Conception or design of the work: all authors. Data collection: all authors. Data analysis and interpretation: RB, GKJH and N'tH. Drafting the article: RB and MJvdV. Critical revision of the article: all authors. Final approval of the version to be published: all authors.

Funding This study was funded by Merck Sharp and Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA.

Competing interests RB: none. N'tH: MSD (unrestricted grant), Pfizer (personal fee). KM: Roche (research grant, personal fee), MSD (research grant, personal fee), Astra Zeneca (research grant, personal fee), Pfizer (personal fee), Benecke (personal fee), BMS (personal fee), Abbvie (personal fee), Diaceutics (personal fee). E-JS: MSD (research grant, personal fee), BMS (research grant, personal fee), Novartis (research grant), AstraZeneca (research grant), AbbVie (personal fee), Bayer (personal fee), Roche (personal fee). ET: HistoGeneX (personal fee), Roche Diagnostics (personal fee). JHVdT: Astellas (research grant), BMS (research grant, personal fee), AbbVie (personal fee), Astra Zeneca (personal fee), Boehringer-Ingelheim (personal fee),

BMS (personal fee), Eli Lilly (personal fee), MSD (personal fee), Pfizer (personal fee), Roche (personal fee). WT: MSD (personal fee), Roche-Ventana (personal fee), Pfizer (personal fee), Astra Zeneca (personal fee), GSK (personal fee), Chiesi (personal fee), Dutch Asthma Fund (research grant), Biotest (personal fee), Novartis (personal fee), Lilly Oncology (personal fee), Boehringer Ingelheim (personal fee). MJvdV: MSD (research grant), Roche (personal fee).

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

ORCID iD
Rogier Butter http://orcid.org/0000-0002-4277-6814

REFERENCES

1 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;373:123–35.
2 Borghaei H, Paz-Ares L, Horn L, *et al*. Nivolumab versus docetaxel in advanced Nonsquamous non-small-cell lung cancer. *N Engl J Med* 2015;373:1627–39.
3 Herbst RS, Baas P, Kim D-W, *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.
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;375:1823–33.
5 Intlekofer AM, Thompson CB. At the bench: preclinical rationale for CTLA-4 and PD-1 blockade as cancer immunotherapy. *J Leukoc Biol* 2013;94:25–39.
6 Tumeh PC, Harview CL, Yearley JH, *et al*. Pd-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* 2014;515:568–71.
7 Fehrenbacher L, Spira A, Ballinger M, *et al*. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (poplar): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet* 2016;387:1837–46.
8 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. *Lancet* 2017;389:255–65.
9 Gandhi L, Rodríguez-Abreu D, Gadgeel S, *et al*. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N Engl J Med* 2018;378:2078–92.
10 Antonia S, Goldberg SB, Balmanoukian A, *et al*. Safety and antitumour activity of durvalumab plus tremelimumab in non-small cell lung cancer: a multicentre, phase 1B study. *Lancet Oncol* 2016;17:299–308.
11 Gulley JL, Rajan A, Spigel DR, *et al*. Avelumab for patients with previously treated metastatic or recurrent non-small-cell lung cancer (javelin solid tumor): dose-expansion cohort of a multicentre, open-label, phase 1B trial. *Lancet Oncol* 2017;18:599–610.
12 Antonia SJ, Villegas A, Daniel D, *et al*. Durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *N Engl J Med* 2017;377:1919–29.
13 Maleki Vareki S, Garrigós C, Duran I. Biomarkers of response to PD-1/PD-L1 inhibition. *Crit Rev Oncol Hematol* 2017;116:116–24.
14 Chowell D, Morris LGT, Grigg CM, *et al*. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* 2018;359:582–7.
15 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.
16 Thunnissen E, Allen TC, Adam J, *et al*. Immunohistochemistry of pulmonary biomarkers: a perspective from members of the pulmonary pathology Society. *Arch Pathol Lab Med* 2018;142:408–19.
17 McLaughlin J, Han G, Schalper KA, *et al*. Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer. *JAMA Oncol* 2016;2:46–54.
18 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.
19 Scheel AH, Dietel M, Heukamp LC, *et al*. Harmonized PD-L1 immunohistochemistry for pulmonary squamous-cell and adenocarcinomas. *Mod Pathol* 2016;29:1165–72.
20 Scheel AH, Baenfer G, Baretton G, *et al*. Interlaboratory concordance of PD-L1 immunohistochemistry for non-small-cell lung cancer. *Histopathology* 2018;72:449–59.
21 Dolled-Filhart M, Roach C, Toland G, *et al*. Development of a companion diagnostic for pembrolizumab in Non–Small cell lung cancer using immunohistochemistry for programmed death ligand-1. *Arch Pathol Lab Med* 2016;140:1243–9.
22 Rebelatto MC, Midha A, Mistry A, *et al*. Development of a programmed cell death ligand-1 immunohistochemical assay validated for analysis of non-small cell lung cancer and head and neck squamous cell carcinoma. *Diagn Pathol* 2016;11:95.
23 Thunnissen E, de Langen AJ, Smit EF. PD-L1 IHC in NSCLC with a global and methodological perspective. *Lung Cancer* 2017;113:102–5.
24 Krippendorff K. Estimating the reliability, systematic error and random error of interval data. *Educ Psychol Meas* 1970;30:61–70.
25 Warrens MJ. Inequalities between multi-rater kappas. *Adv Data Anal Classif* 2010;4:271–86.
26 Vennapusa B, Baker B, Kowanetz M, *et al*. Development of a PD-L1 complementary diagnostic immunohistochemistry assay (SP142) for Atezolizumab. *Appl Immunohistochem Mol Morphol* 2019;27:92–100.
27 Cooper WA, Tran T, Vilain RE, *et al*. Pd-L1 expression is a favorable prognostic factor in early stage non-small cell carcinoma. *Lung Cancer* 2015;89:181–8.
28 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.
29 Thunnissen E. How to validate predictive immunohistochemistry testing in pathology? A practical approach exploiting the heterogeneity of programmed death ligand-1 present in non-small cell lung cancer. *Arch Pathol Lab Med* 2019;143:11–12.
30 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.
31 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;30:1411–21.
32 Ilie M, Juco J, Huang L, *et al*. Use of the 22C3 anti-programmed death-ligand 1 antibody to determine programmed death-ligand 1 expression in cytology samples obtained from non-small cell lung cancer patients. *Cancer Cytopathol* 2018;126:264–74.
33 Russell-Goldman E, Kravets S, Dahlberg SE, *et al*. Cytologic-histologic correlation of programmed death-ligand 1 immunohistochemistry in lung carcinomas. *Cancer Cytopathol* 2018;126:253–63.
34 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;13:1302–11.
35 Munari E, Rossi G, Zamboni G, *et al*. Pd-L1 assays 22C3 and SP263 are not interchangeable in Non–Small cell lung cancer when considering clinically relevant cutoffs. *Am J Surg Pathol* 2018;42:1384–9.
36 Sheffield BS, Fulton R, Kalloger SE, *et al*. Investigation of PD-L1 biomarker testing methods for PD-1 axis inhibition in non-squamous non-small cell lung cancer. *J Histochem Cytochem* 2016;64:587–600.
37 Kim H, Kwon HJ, Park SY, *et al*. PD-L1 immunohistochemical assays for assessment of therapeutic strategies involving immune checkpoint inhibitors in non-small cell lung cancer: a comparative study. *Oncotarget* 2017;8:98524–32.
38 Xu H, Lin G, Huang C, *et al*. Assessment of concordance between 22C3 and SP142 immunohistochemistry assays regarding PD-L1 expression in non-small cell lung cancer. *Sci Rep* 2017;7:16956.
39 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;29:953–8.
40 Pang C, Yin L, Zhou X, *et al*. Assessment of programmed cell death ligand-1 expression with multiple immunohistochemistry antibody clones in non-small cell lung cancer. *J Thorac Dis* 2018;10:816–24.
41 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;23:3585–91.
42 Soo RA, Yun Lim JS, Asuncion BR, *et al*. Determinants of variability of five programmed death ligand-1 immunohistochemistry assays in non-small cell lung cancer samples. *Oncotarget* 2018;9:6841–51.