Clinical Cancer Research

Intra- and Interobserver Reproducibility Assessment of PD-L1 Biomarker in Non–Small Cell Lung Cancer

Wendy A. Cooper[1,2,3], Prudence A. Russell[4], Maya Cherian[5], Edwina E. Duhig[6], David Godbolt[7], Peter J. Jessup[8], Christine Khoo[9], Connull Leslie[10], Annabelle Mahar[11], David F. Moffat[12], Vanathi Sivasubramaniam[13], Celine Faure[14], Alena Reznichenko[15], Amanda Grattan[15], and Stephen B. Fox[9]

Abstract

Purpose: Reliable and reproducible methods for identifying PD-L1 expression on tumor cells are necessary to identify responders to anti–PD-1 therapy. We tested the reproducibility of the assessment of PD-L1 expression in non–small cell lung cancer (NSCLC) tissue samples by pathologists.

Experimental Design: NSCLC samples were stained with PD-L1 22C3 pharmDx kit using the Dako Autostainer Link 48 Platform. Two sample sets of 60 samples each were designed to assess inter- and intraobserver reproducibility considering two cut points for positivity: 1% or 50% of PD-L1 stained tumor cells. A randomization process was used to obtain equal distribution of PD-L1 positive and negative samples within each sample set. Ten pathologists were randomly assigned to two subgroups. Subgroup 1 analyzed all samples on two consecutive days. Subgroup 2 performed the same assessments, except they received a 1-hour training session prior to the second assessment.

Results: For intraobserver reproducibility, the overall percent agreement (OPA) was 89.7% [95% confidence interval (CI), 85.7–92.6] for the 1% cut point and 91.3% (95% CI, 87.6–94.0) for the 50% cut point. For interobserver reproducibility, OPA was 84.2% (95% CI, 82.8–85.5) for the 1% cut point and 81.9% (95% CI, 80.4–83.3) for the 50% cut point, and Cohen's κ coefficients were 0.68 (95% CI, 0.65–0.71) and 0.58 (95% CI, 0.55–0.62), respectively. The training was found to have no or very little impact on intra- or interobserver reproducibility.

Conclusions: Pathologists reported good reproducibility at both 1% and 50% cut points. More adapted training could potentially increase reliability, in particular for samples with PD-L1 proportion, scores around 50%. *Clin Cancer Res; 23(16); 4569–77. ©2017 AACR.*

Introduction

Lung cancer is the leading cause of cancer-related death worldwide (1, 2). Approximately 80% of patients with newly diagnosed non–small cell lung cancer (NSCLC) present with inoperable locally advanced or metastatic disease (3). Platinum-based chemotherapy, with or without maintenance therapy and subsequently followed by second-line cytotoxic chemotherapy, is standard treatment for most patients with advanced NSCLC, with a median survival of approximately 1 year (4, 5). Despite the development of targeted therapies for selected patient subgroups, the majority of patients do not attain prolonged disease control, and in the United States, the 5-year survival rates vary greatly depending on cancer stage, from 1% for patients with Stage IV up to 49% for patients with Stage IA NSCLC (http://www.cancer.org/acs/groups/cid/documents/webcontent/003115-pdf.pdf), emphasizing the urgent need for more effective therapies.

Programmed-death 1 (PD-1 or CD279) is a type I trans-membrane protein that is an important inhibitory co-receptor expressed by T cells, B cells, monocytes, natural killer cells, dendritic cells and many tumor-infiltrating lymphocytes (6). Programmed-death ligand 1 (PD-L1 also known as B7-H1 or CD274) is a ligand for the PD-1 receptor and is a key-immune-checkpoint receptor expressed by activated T cells. Interaction of PD-1 with its ligand PD-L1 leads to apoptosis or inactivation of activated T cells. PD-L1 expression can be upregulated in tumor

[1]Tissue Pathology and Diagnostic, Oncology, Royal Prince Alfred Hospital, New South Wales, Australia. [2]Sydney Medical School, The University of Sydney, Sydney, Australia. [3]School of Medicine, Western Sydney University, Sydney, Australia. [4]Department of Anatomical Pathology, St Vincent's, Hospital and University of Melbourne, Victoria, Australia. [5]The Canberra Hospital, Garran, Australian Capital Territory, Australia. [6]Sullivan Nicholaides Pathology, Tugun Lab, C/o John Flynn Hospital, Queensland, Australia. [7]Pathology Queensland—The Prince Charles Laboratory, The Prince Charles Hospital, Queensland, Australia. [8]Royal Hobart Hospital Pathology Service, Hobart, Tasmania, Australia. [9]Department of Pathology, Peter MacCallum Cancer Centre and University of Melbourne, Victoria, Australia. [10]Department of Anatomical Pathology, PathWest Laboratory Medicine, QEII Medical Centre, Western Australia, Australia. [11]Royal Prince Alfred Hospital, Department of Tissue Path and Diagnostic Oncology, Camperdown, New South Wales, Australia. [12]SA Pathology, Department of Anatomical Pathology, FMC, Bedford Park, South Australia, Australia. [13]SydPath St Vincents Hospital, Department of Anatomical Pathology, Darlinghurst, New South Wales, Australia. [14]Mapi Group, Real World Evidence, Villette, Lyon, France. [15]MSD (Australia), Macquarie Park, New South Wales, Australia.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Wendy A. Cooper, Royal Prince Alfred Hospital, Missenden Road, Camperdown, NSW 2050, Australia. Phone: 61-2-9515-7458; Fax: 61-2-9515-8405. E-mail: Wendy.Cooper@sswahs.nsw.gov.au

doi: 10.1158/1078-0432.CCR-17-0151

Translational Relevance

Reliable and reproducible methods for detecting PD-L1 expression on tumor cells are necessary for identifying patients who respond to anti–PD-1 therapy. The PD-L1 IHC 22C3 pharmDx is a companion diagnostic developed by Dako to identify patients with non–small cell lung cancer who are likely to derive benefit from treatment with pembrolizumab. Using this assay, we investigated the inter- and intraobserver reproducibility of pathologists' assessment of PD-L1 staining and the impact of training on reproducibility. Our findings show good agreement with both 1% and 50% cut points. A one-hour training session was found to have no or very little impact on the inter- or intraobserver reproducibility and to slightly improve agreement with gold standard PD-L1 tumor proportion score for samples with a tumor proportion score >40%.

cells as an adaptive mechanism to avoid host immunity, or intrinsically as a result of oncogene activation, or loss of tumor suppressor genes, such as PTEN (7). Consequently, PD-L1 is frequently expressed in a wide range of human tumors, including NSCLC (8–13). Monoclonal antibodies blocking the checkpoints PD-1/PD-L1 have shown great promise in treating many diverse cancer types, with durable responses in some patients with advanced disease, including NSCLC (8, 13–22). Early data from multinational clinical trials found that approximately 61% of patients with advanced NSCLC are PD-L1 positive, that is, ≥1% PD-L1 expression in tumor cells, and that approximately 23% of patients are strongly positive, that is, ≥50% PD-L1 expression in tumor cells (20). In an effort to predict patients who will respond to checkpoint blockade, cancers have been categorized into four different tumor environments based on their PD-L1 status and presence or absence of tumor-infiltrating lymphocytes (TIL): Type I—PD-L1 positive with TILs driving adaptive immune resistance; Type II—PD-L1 negative with no TILs indicating immunologic ignorance; Type III—PD-L1 positive with no TILs indicating intrinsic induction; and Type IV—PD-L1 negative with TILs indicating immune tolerance (7).

Most evidence shows that PD-L1 expression levels in tumor cells correlate with increased response to anti-PD-1 therapies in NSCLC patients treated with nivolumab (19, 23), atezolizumab (24), and pembrolizumab (20, 25). The cut points for a positive PD-L1 expression are not homogeneous in the literature, ranging from ≥1% to ≥50% depending on the PD-L1 antibody used (18, 20, 25). This lack of homogeneity results in a wide variability in the reported range of percentage of tumor samples expressing PD-L1 (21%–95%; ref. 18) and in a possible lack of clarity when assessing responses to treatment (20, 25). In patients with NSCLC treated with pembrolizumab, progression-free and overall survival (OS) were shorter among patients with a PD-L1 (22C3 clone) tumor proportion score of 1% to 49% or among patients with a score of less than 1% compared with those with a score of at least 50% (13, 20). Another study in patients with NSCLC treated with second line pembrolizumab demonstrated that although there was no difference in OS between patients who had no expression of PD-L1 versus those with any staining (≥1%), patients whose tumors exhibited high expression of PD-L1 (≥50%) were more likely to have a longer OS, independent of patient age and tumor stage (20). In a recent clinical trial in patients with advanced NSCLC with a tumor PD-L1 proportion score of at least 50%, first-line treatment with pembrolizumab was associated with significantly longer progression-free survival and OS, with fewer adverse events than the investigators' choice platinum-based chemotherapy (13). In the sub-analysis of a randomized, open-label phase II study investigating the efficacy of pembrolizumab combined with first-line chemotherapy in patients with non-squamous NSCLC irrespective of PD-L1 expression, objective response was achieved by 26% of patients with 1%-49% PD-L1 tumor proportion score and by 80% of patients with a PD-L1 tumor proportion score ≥50% in the active arm (22).

The availability of high-quality-biomarker assays is critical in guiding clinical practice (26). Currently, there is a little data examining the robustness of the PD-L1 assay. Thus, the primary objective of the study was to test the intra- and interobserver reproducibility of pathologists' scoring of PD-L1 in NSCLC using the FDA-approved companion diagnostic PD-L1 22C3 PharmDx Kit (Dako North America Inc.) at the two cut points generally accepted for this antibody (i.e., 1% and 50%; ref. 13, 20, 22). The study also aimed to assess the impact of a one-hour training on reproducibility of pathologists' scoring.

Materials and Methods

Tissue samples and immunohistochemistry

De-identified formalin-fixed paraffin-embedded tissue microarrays with 1-mm cores of surgically resected early stage NSCLC samples were obtained from Peter MacCallum Cancer Centre (Melbourne), St Vincent's Hospital (Melbourne), and Royal Prince Alfred Hospital (Sydney). Selected tumor samples derived from a unique patient, contained more than 100 cells per sample and were all less than 15 years old. Approval for this study was obtained from the Ethics Review Committees of the relevant institutions. Samples were stained for PD-L1 at the Peter MacCallum Cancer Centre with the 22C3 pharmDx Kit and the Dako Autostainer Link 48 platform. The deparaffinization, rehydration, and target retrieval procedure were performed using the EnVision FLEX Target Retrieval solution (low pH, 1×) and the EnVision FLEX wash buffer 1×. After this 3 in 1 procedure, the tissue samples were placed on Autostainer Link 48. The instrument performed the staining process by applying the appropriate reagent, monitoring the incubation time and rinsing slides between reagents. The reagent times were preprogrammed in the Dako Link software. Omission of the primary antibody was used as a negative control. Tissue samples were subsequently counterstained with hematoxylin and mounted in non-aqueous, permanent mounting media.

Immunohistochemistry assessment

PD-L1 expression was defined as the percentage of viable tumor cells with any perceptible membrane staining irrespective of staining intensity to determine the "tumor proportion score". Thus, pathologists were instructed to include any partial or complete membrane staining (intensity ≥+1, i.e., weak staining) that was perceived as being distinct from cytoplasmic staining in the scoring. Normal cells and tumor-associated immune cells were excluded from the scoring. Tumor samples stained with the negative control reagent must have 0 specific staining and <1+ background staining to be considered as acceptable.

Gold standard PD-L1 tumor proportion score

A gold standard PD-L1 tumor proportion score was established for all the tissue samples meeting the eligibility criteria. Two lead investigators were trained and certified to assess PD-L1 22C3 pharmDx staining undergoing a 2-day training course by Dako prior to the experiment. Each of the two lead investigators assessed all available samples for PD-L1 staining, histopathology, and background quality, and the associated negative controls. For all assessable samples, lead investigators assessed the percentage of PD-L1 positive tumor cells. The gold standard was established as a consensus using a multi-headed microscope. When a consensus could not be reached the case was excluded. Using the gold standard PD-L1 tumor proportion scores, an independent statistician was able to perform a stratified randomization for each sample set.

Number of assessments necessary to evaluate inter- and intraobserver reproducibility

Considering an expected true overall percent agreement (OPA) $\geq$ 89% and a two-tailed α of 5%, it was estimated that a total number of 300 pairwise comparisons would be necessary to meet the study acceptance criterion: lower bound of the Wilson 95% confidence interval (CI) of the OPA $\geq$85%.

Two separate sample sets were generated to assess intra- and inter-reproducibility for the two cut points, 1% and 50%. A randomization process was used to constitute the two sample sets to avoid biases that could influence the evaluation of reproducibility. Both sample sets were designed to contain equally distributed PD-L1 positive and negative samples; and one-third of the samples in the sample sets were negative and positive tissue samples around the cut point. For the 1% cut point sample set, 30 negative samples (<1%); 10 positive samples close to cut point ([1%–10%]); and 20 positive samples far from the cut point (>10%) were randomly selected from the gold standard cases. For the 50% cut point sample set, 20 negative samples far from the cut point (<35%); 10 negative samples close to the cut point ([35%–49%]); 10 positive samples close to the cut point ([50%–60%]); and 20 positive samples far from the cut point (>60%) were randomly selected.

Assessment of PD-L1 staining by the participating pathologists

The study experiment was conducted on two separate days (Fig. 1). On the first day, participating pathologists (PP) assessed the PD-L1 staining without any specific training but with written instructions to assess the tumor proportion score for each sample as outlined above. The comparison between the PP's assessments of each sample from the sample set provided data on the interobserver reproducibility. For the second day, PPs were randomly assigned in two subgroups: Subgroup 1—PPs assessed all the samples for a second time. Comparison with their previous assessments provided data on intraobserver reproducibility; Subgroup 2—PPs were trained on scoring PD-L1 immunohistochemistry before assessing the samples for the second time. Training consisted of a one-hour presentation covering the biology of PD-L1, development of the assay, cellular expression, and strategies to optimally assess expression in NSCLCs. Comparison with their first assessment provided data on the impact of training on their assessment. For both subgroups, samples were not assessed in the same order during the two assessments to avoid recall bias.

Participating pathologists

To obtain good representation of both samples and pathologists, it was decided that a minimum of 5 PPs in each subgroup were needed to assess the 60 samples per sample set to meet the necessary 300 pairwise comparisons. Ten PPs were selected from across all states in Australia to represent a range of pathologists' experience and type of practice and included 8 PPs working in public laboratories, 1 in a private laboratory and 1 in a mixed laboratory. PPs' median age was 46 years old (range: 35–68 years) with a median 15 years of experience (range: 5.0–19.0 years). PPs were presented with a single "sample set" for assessment, resulting from pooling the 1% cut point and the 50% cut point sample sets together.

Statistical analysis

Statistical analyses were undertaken using SAS software version 9.2. Using percentage of staining provided by PPs, samples were

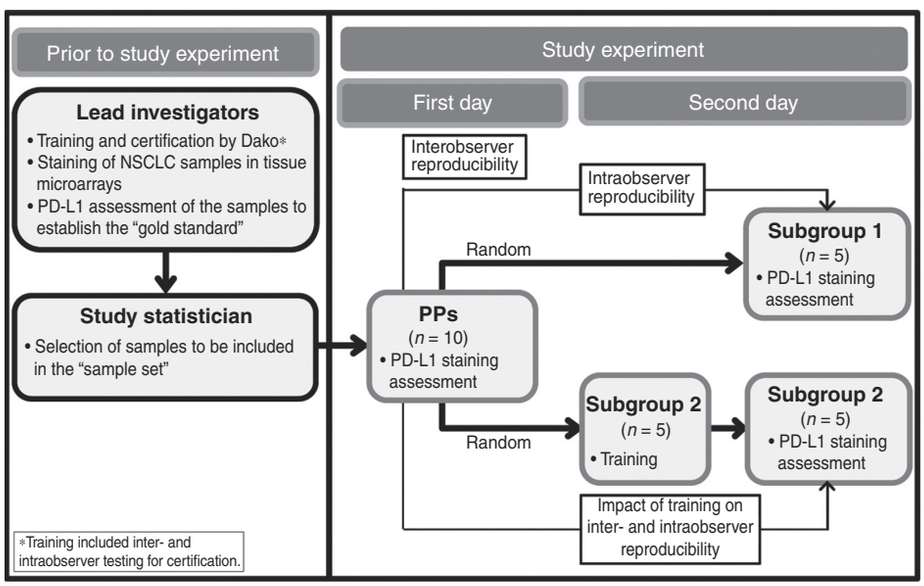

Figure 1.
Study design. On the first day, the 10 PPs assessed PD-L1 tumor proportion score for each of the 60 samples in the sample set. The comparison between the PP's assessments of each provided data on the interobserver reproducibility. For the second day, PPs were randomly assigned in two subgroups of five PPs: PPs in subgroup 1 assessed all the samples for a second time. Comparison with their previous assessments provided data on intraobserver reproducibility; PPs in subgroup 2 were trained on scoring PD-L1 immunohistochemistry before assessing the samples for the second time. Comparison with their first assessment provided data on the impact of training on their assessment. PP, participating pathologist.

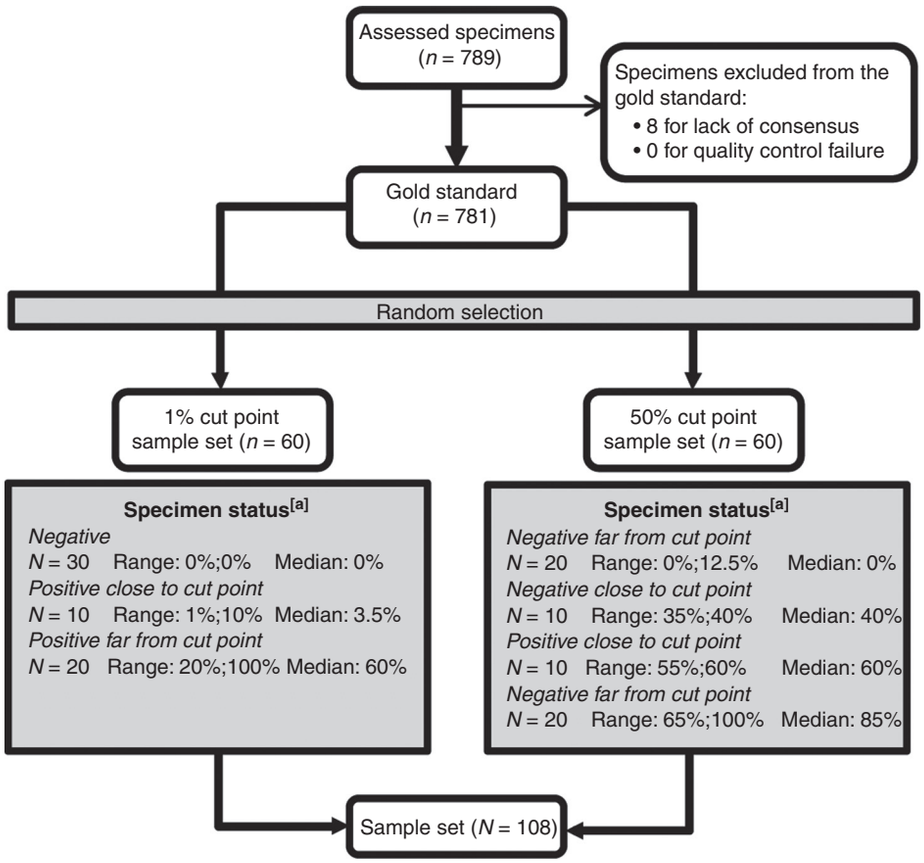

Figure 2.
Disposition of the samples. Two separate sample sets of 60 samples were randomly generated from the gold standard to assess intra- and inter-reproducibility for the two cut points, 1% and 50%. Both sample sets were designed to contain equally distributed PD-L1 positive and negative samples; and one third of the samples in the sample sets were negative and positive tissue samples around the cut point. Participating pathologists were presented with a single "sample set" of 108 samples for assessment, resulting from pooling the 1% cut point and the 50% cut-point sample sets together.

[a]According to gold standard.

adjudicated as positive or negative according to cut points, 1% and 50%. The inter- and intraobserver reproducibility was assessed using OPA, positive and negative percent agreement (PPA and NPA), and Cohen's κ coefficient, whenever applicable. Agreement of PPs' assessment with the gold standard was assessed using sensitivity (or true positive rate), specificity (or true negative rate), positive and negative predictive values. Ninety-five percent of CIs were computed for all measurements.

Bias-adjusted κ (BAK) and prevalence-adjusted bias-adjusted κ (PABAK) were calculated to assess the presence of index bias (index bias being the extent to which the observers disagree on the proportion of positive or negative cases) and prevalence bias (bias related to the distribution of positive and negative specimens in the analyzed sample set) influencing Cohen's κ range.

Results

Disposition of the samples

Among the 789 samples assessed to establish the gold standard PD-L1 tumor proportion score, 8 were discarded for lack of consensus and none due to quality control failure (Fig. 2). Among the 781 remaining samples, 75.7% had a PD-L1 tumor proportion score <1%; 14.0% had a PD-L1 tumor proportion score between 1% and 49%; and 10.3% had a PD-L1 tumor proportion score ≥50%. Examples of a negative sample, a positive sample at 1% cut point and a positive sample at 50% cut point are presented in Fig. 3 (concordant cases, A–C). As 12 samples were included in both 1% and 50% sample sets, 108 samples were included in the sample set provided to the PPs.

Intraobserver reproducibility

Samples were assessed twice by PPs in subgroup 1, resulting in five pairwise comparisons per sample. As sample sets included 60 samples, there were 300 pairwise comparisons for the 1% and the 50% cut point sample sets (Table 1). For the 1% cut point sample set, 269 of the pairwise comparisons were concordant, giving an OPA of 89.7% (95% CI, 85.7–92.6). For 87.1% of the 31 non-concordant pairwise comparisons the two assessments differed by less than 10%. For the 50% cut point sample set, 274 of the pairwise comparisons were concordant, giving an OPA of 91.3 (95% CI, 87.6–94.0). For 42.3% of the 26 non-concordant pairwise comparisons the two assessments differed by less than 10%.

Interobserver reproducibility

Each sample assessment (positive/negative) was compared with the assessments for the same sample made by the nine other PPs, resulting in 45 pairwise comparisons per sample. As both sample sets included 60 samples, there were 2,700 pairwise comparisons for each sample set (Table 2). For the 1% cut point sample set, 2,273 of the pairwise comparisons were concordant, giving an OPA of 84.2% (95% CI, 82.8–85.5). For 80.6% of the 427 non-concordant pairwise comparisons the two assessments were reported to be within the range 0%–10%. For the 50% cut point sample set, 2,211 of the pairwise comparisons were

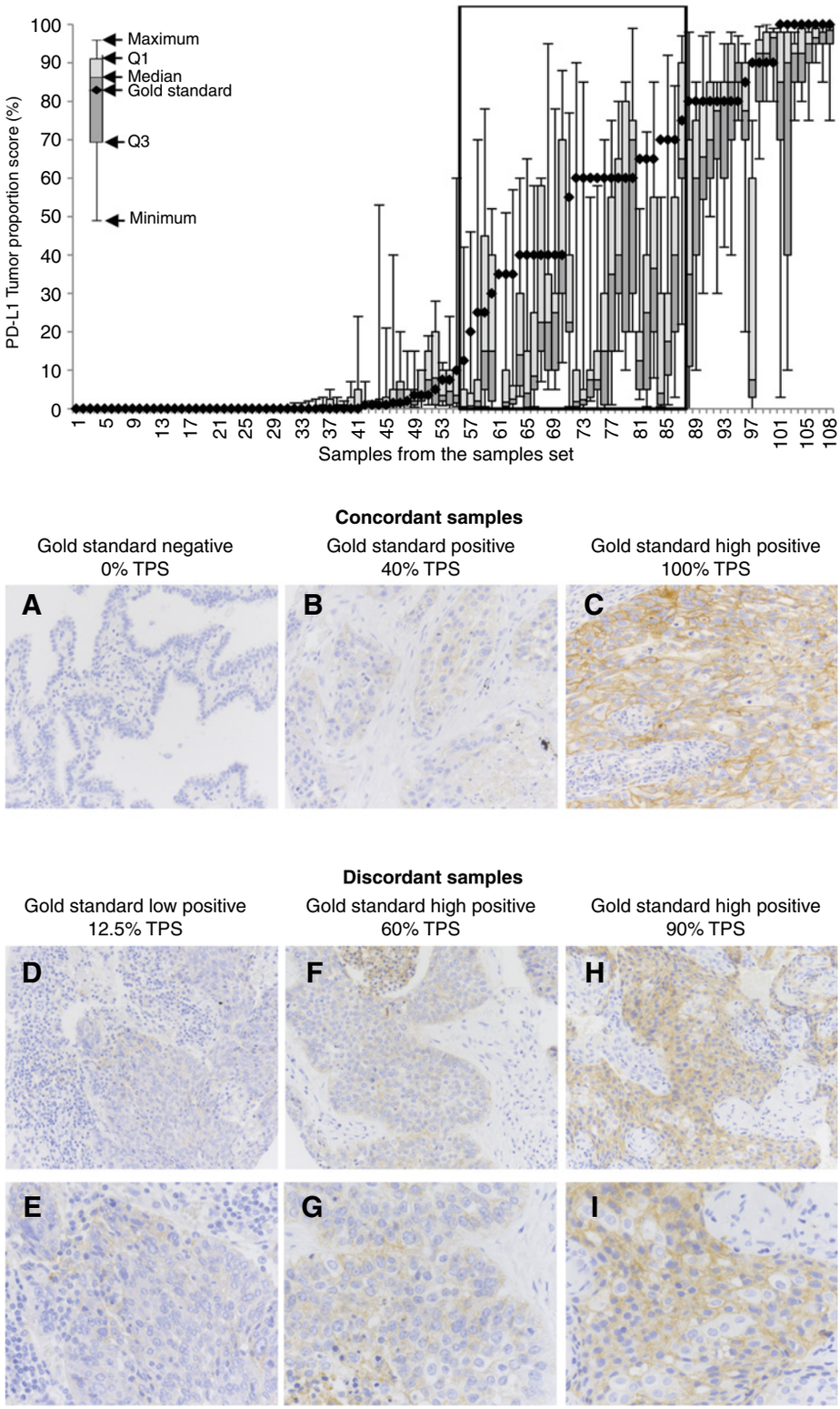

Figure 3.

Variability in participating pathologist's assessment of the PD-L1 tumor proportion score. Abbreviations: TPS, tumor proportion score; PP, participating pathologist. Box plot. The box plot represents the variability in PPs' assessment of the PD-L1 TPS for each sample. Based on box plot representation, photomicrographs of samples for which PPs' assessment was concordant or discordant with gold standard are presented underneath. Concordant Samples. (Magnification ×200). **A,** Sample 1: PD-L1 negative by gold standard with a 0% PD-L1 TPS. According to PPs, median PD-L1 TPS was 0% (range 0%–0%). **B,** Sample 66: PD-L1 positive by gold standard with a 40% PD-L1 TPS. According to PPs, median PD-L1 TPS was 8.5% (interquartile range: 5%–25%). **C,** Sample 105: PD-L1 positive by gold standard with a 100% PD-L1 TPS. According to PPs, median PD-L1 TPS was 95% (range 75%–100%, interquartile range 90%–98%). Discordant Samples. Left panel (**D** and **E**). Sample 56: PD-L1 low positive by gold standard assessment with a 12.5% PD-L1 TPS. According to PPs, median PD-L1 TPS was 0.5, i.e., most PPs scored <1% TPS. **D,** At medium power (×200), the tumor appears to be negative, but (**E**) high power (×400) shows focal weak membranous staining. Middle panel (**F** and **G**). Sample 73: PD-L1 high positive by gold standard assessment with a 60% PD-L1 TPS. According to PPs, median PD-L1 TPS was 2.35% (range 0%–85%), i.e., most PPs scored <50% TPS. **F,** At medium power (×200), the tumor is difficult to score, whereas (**G**) high power (×400) shows weak membranous staining in the majority of tumor cells. Right panel (**H** and **I**). Sample 97: PD-L1 high positive by gold standard assessment with a 90% PD-L1 TPS. According to PPs, median PD-L1 TPS was 7.5% (range 0%–75%), i.e., most PPs scored <50% PD-L1 TPS. **H,** At medium power, the tumor shows membranous staining that is difficult to distinguish from cytoplasmic staining (×200), whereas (**I**) high power (×400) shows membranous staining in the majority of tumor cells that is distinct from the cytoplasmic staining.

concordant, giving an OPA of 81.9 (95% CI, 80.4–83.3). For 4.3% of the 489 non-concordant pairwise comparisons were reported to be within the range 40%–60%. Compared with the 1% cut point sample, the percentage of non-concordant pairwise comparison with the two assessments close to one-another (<10%) was low for the 50% cut point. This discrepancy resulted from the

Table 1. Intraobserver reproducibility

	Intraobserver reproducibility[a]	
	1% cut point (*N* = 300)	50% cut point (*N* = 300)
Pairwise comparison similarly evaluated		
No	31 (10.3%)	26 (8.7%)
Yes	269 (89.7%)	274 (91.3%)
Results of pairwise comparisons[b]		
Negative–Negative	145 (48.3%)	198 (66.0%)
Negative–Positive	12 (4.0%)	13 (4.3%)
Positive–Negative	19 (6.3%)	13 (4.3%)
Positive–Positive	124 (41.3%)	76 (25.3%)
Measures of agreement		
OPA (%) [95% CI]	89.7 (85.7–92.6)	91.3 (87.6–94.0)

Abbreviations: OPA, overall percent agreement; CI, confidence interval.
[a]The 300 pairwise comparisons are defined based on the 60 samples evaluated and five unique pairwise comparisons per sample for the five participating pathologists in subgroup 1.
[b]Results are given in the order: First assessment–Second assessment.

difference in the number of samples available within the 10% range around the cut point: 40 samples from the gold standard close to the 1% cut point, compared with 16 close to the 50% cut point.

Cohen's κ coefficient was 0.68 (95% CI, 0.65–0.71) for the 1% cut point sample set, indicating substantial interobserver agreement (27), and 0.58 (95% CI, 0.55–0.62) for the 50% cut point sample set, indicating moderate interobserver agreement. For the 1% cut point sample sets, BAK and PABAK were similar to Cohen's κ coefficient, indicating that no bias influenced the Cohen's κ magnitude. However, for the 50% cut point sample set, PABAK was higher than the Cohen's κ coefficient [0.64 (95% CI, 0.61–0.67) vs. 0.58 (95% CI, 0.55–0.62)], indicating a prevalence bias in PPs' assessments. Approximately 74% of the samples in the 50% cut point sample were considered PD-L1 negative by at least half of the 10 PPs and approximately 26% were considered PD-L1 positive.

Table 2. Interobserver reproducibility

	Interobserver reproducibility[a]	
	1% cut point (*N* = 2,700)	50% cut point (*N* = 2,700)
Pair-wise comparison similarly evaluated		
No	427 (15.8%)	489 (18.1%)
Yes	2,273 (84.2%)	2,211 (81.9%)
Results of pair-wise comparisons[b]		
Negative–Negative	1,231 (45.6%)	1,614 (59.8%)
Negative–Positive	217 (8.0%)	380 (14.1%)
Positive–Negative	210 (7.8%)	109 (4.0%)
Positive–Positive	1,042 (38.6%)	597 (22.1%)
Measures of agreement (95% CI)		
OPA (%)	84.2 (82.8–85.5)	81.9 (80.4–83.3)
NPA (%)	85.0 (83.1–86.8)	80.9 (79.2–82.6)
PPA (%)	83.2 (81.1–85.2)	84.6 (81.7–87.0)
Cohen's κ coefficient	0.68 (0.65–0.71)	0.58 (0.55–0.62)
BAK	0.68 (0.65–0.71)	0.58 (0.54–0.61)
PABAK	0.68 (0.66–0.71)	0.64 (0.61–0.67)

NOTE: Cohen's κ coefficient: κ value range from −1 to +1, with −1 indicating perfect disagreement and +1 indicating perfect agreement between the pathologists. The strength of this agreement is defined as poor if $\kappa < 0.00$, slight if κ was within 0.00–0.20, fair if κ was within 0.21–0.40, moderate if κ was within 0.41–0.60, substantial if κ was within 0.61–0.80, and almost perfect if κ was within 0.81–1.00.
Abbreviations: OPA, overall percent agreement; NPA, negative percent agreement; PPA, positive percent agreement; CI, confidence interval; BAK, bias-adjusted κ; PABAK, prevalence-adjusted bias-adjusted κ.
[a]The 2700 pair-wise comparisons were defined based on 60 samples evaluated and 45 unique pair-wise comparisons per sample for the 10 participating pathologists.
[b]The first status Negative or Positive corresponds to the median status among the 10 assessments by the 10 participating pathologists.

Impact of training on interobserver reproducibility

To assess the impact of training on interobserver reproducibility, each sample assessment (positive/negative) was compared with the assessments for the same sample made by the four other PPs, resulting in 10 pairwise comparisons per sample. Assessments performed before and after training were compared. As sample sets included 60 samples, there were 600 pairwise comparisons for each sample set (Table 3). For the 1% cut point sample set, there were 492 and 494 concordant assessments for the first and second assessments, respectively. The OPAs were similar for the first and second assessments, 82.0 (95% CI, 78.7–84.9) and 82.3 (95% CI, 79.1–85.2), respectively, meaning that the training had no impact on the interobserver reproducibility at 1%. For the 50% cut point sample set, there were 470 and 490 concordant assessments for the first and second assessments, respectively. The OPAs were similar for the first and second assessments, 78.3 (95% CI, 74.9–81.4) and 81.7 (95% CI, 78.4–84.6), respectively, showing a slight improvement for the second assessment, meaning that the training had little impact on the interobserver reproducibility at 50%.

Agreement between pathologists' assessments and the gold standard

To assess agreement with gold standard, each sample assessment from the 10 PPs was compared with the gold standard. For the 1% sample set, 87.8% of assessments were concordant, resulting in a 84.3% (95% CI, 80.2–88.5) sensitivity (probability that a pathologist correctly scored the specimen as positive among the positive specimens) and a 91.3% (95% CI, 88.2–94.5) specificity (probability that a pathologist correctly scored the specimen as negative among the negative specimens). The negative and positive predictive values were 85.4% (95% CI, 81.5–89.2) and 90.7% (95% CI, 87.3–94.1), respectively. For the 50% evaluation sample set, 75.2% of assessments were concordant, resulting in a 56.3% (95% CI, 50.7–62.0) sensitivity and a 94.0% (95% CI, 91.3–96.7) specificity. The negative and positive predictive values were 68.3% (95% CI, 63.8–72.8) and 90.4% (95% CI, 86.2–94.6), respectively, suggesting that the PPs had difficulties in correctly assessing the positive specimens. Agreement of each individual PP with the gold standard PD-L1 tumor proportion score was also assessed and results once again highlighted PPs difficulty when assessing positivity at the 50% cut point (Supplementary Table S1). Median PPs' agreement was 89.2% (range, 78.3%–93.3%) for the 1% cut point sample set and 74.2% (range, 66.7%–85.0%) for the 50% sample-set.

Box plots were used to assess the variability in PPs' assessments for each sample (Fig. 3). Variability in PPs assessment was high for samples with PD-L1 tumor proportion score between 30% and 80% (gold standard assessment score). Compared with the gold standard, PPs had a tendency to underestimate the PD-L1 tumor proportion score, although several PPs also overestimated PD-L1 tumor proportion score. Review of samples where PPs median PD-L1 tumor proportion score was incorrect compared with the gold standard (i.e., PPs median PD-L1 tumor proportion score

Table 3. Impact of training on the inter observer reproducibility

	1% Cut point ($N = 600$)[a]		50% Cut point ($N = 600$)[a]	
	First assessment	Second assessment	First assessment	Second assessment
Pair-wise comparison similarly evaluated				
No	108 (18.0%)	106 (17.7%)	130 (21.7%)	110 (18.3%)
Yes	492 (82.0%)	494 (82.3%)	470 (78.3%)	490 (81.7%)
Results of pair-wise comparisons[b]				
Negative–Negative	274 (45.7%)	277 (46.2%)	339 (56.5%)	329 (54.8%)
Negative–Positive	46 (7.7%)	46 (7.7%)	108 (18.0%)	78 (13.0%)
Positive–Negative	62 (10.3%)	60 (10.0%)	22 (3.7%)	32 (5.3%)
Positive–Positive	218 (36.3%)	217 (36.2%)	131 (21.8%)	161 (26.8%)
Measures of agreement (95% CI)				
OPA (%)	82.0 (78.7–84.9)	82.3 (79.1–85.2)	78.3 (74.9–81.4)	81.7 (78.4–84.6)
NPA (%)	85.6 (81.4–89.0)	85.8 (81.5–89.2)	75.8 (71.7–79.6)	80.8 (76.7–84.4)
PPA (%)	77.9 (72.6–82.3)	78.3 (73.1–82.8)	85.6 (79.2–90.3)	83.4 (77.5–88.0)
Cohen's κ coefficient	0.64 (0.58–0.70)	0.64 (0.58–0.70)	0.52 (0.45–0.59)	0.60 (0.54–0.67)
BAK	0.64 (0.57–0.70)	0.64 (0.58–0.70)	0.51 (0.43–0.58)	0.60 (0.53–0.67)
PABAK	0.64 (0.58–0.70)	0.65 (0.59–0.71)	0.58 (0.51–0.64)	0.64 (0.57–0.70)

NOTE: Cohen's κ coefficient: κ value range from −1 to +1, with −1 indicating perfect disagreement and +1 indicating perfect agreement between the pathologists. The strength of this agreement is defined as poor if $\kappa < 0.00$, slight if κ was within 0.00–0.20, fair if κ was within 0.21–0.40, moderate if κ was within 0.41–0.60, substantial if κ was within 0.61–0.80, and almost perfect if κ was within 0.81–1.00.

Abbreviations: OPA, overall percent agreement; NPA, negative percent agreement; PPA, positive percent agreement; CI, confidence interval; BAK, bias-adjusted κ; PABAK, prevalence-adjusted bias-adjusted κ.

[a]The 600 pair-wise comparisons were defined based on 60 samples evaluated and 10 unique pair-wise comparisons per sample for the five participating pathologists in subgroup 2.

[b]The first status Negative or Positive corresponds to the median status among the 10 assessments by the 10 participating pathologists.

was >1% when gold standard was >50%) revealed weak membranous staining or concomitant cytoplasmic staining as a possible cause for PP underscoring (Fig. 3, Concordant and Discordant Samples).

The impact of training on the agreement with the gold standard was assessed. For the 1% sample set, there were 87.3% and 87.7% concordant assessments before and after training, respectively. For the 50% sample set, there were 75.3% and 78.7% concordant assessments before and after training, respectively. No differences in agreement with the gold standard were observed at the 50% cut point in untrained PPs on the first and second assessments, regardless of the PD-L1 tumor proportion score (Supplementary Table S2). In contrast, training slightly improved agreement with gold standard for samples with a PD-L1 tumor proportion score >40% (Supplementary Table S2). Impact was the highest for samples with a PD-L1 tumor proportion score >80% with 86.2% and 95.4% concordant assessments before and after training. These results show that the training received by the five PPs in subgroup 2 had very little impact on the concordance, and that training impact was mainly observed for samples with a very high PD-L1 tumor proportion score (>80%).

Discussion

We investigated the reproducibility of pathologists' assessment of PD-L1 staining in NSCLC samples and observed that a high intra- and interobserver agreement on the scores for NSCLC samples can be obtained with a 1% and 50% cut point, with a greater reproducibility at the 1% cut point. There is limited data examining the reproducibility of assessment of PD-L1 staining in tissue samples from patients with NSCLC and most studies only involve small numbers of pathologists or small numbers of tumors, making it easier to achieve concordance (25, 28, 29). Phillips and colleagues (28), assessed intra- and interobserver reproducibility using the 28-8 PD-L1 antibody and considering 1% and 5% cut points. Three observers assessed 90 samples for intraobserver reproducibility and 270 samples for interobserver reproducibility. They reported good reproducibility of PD-L1 staining assessment with an OPA >90% for all comparisons and lower bounds of the 95% CI >85% with fewer pathologists than our study. Cooper and colleagues (25) assessed the reproducibility of PD-L1 staining in patients with early stage NSCLC using the 22C3 PD-L1 antibody and a 50% cut point. Two pathologists assessed 681 samples and obtained substantial concordance for determining PD-L1 positivity (Cohen's $\kappa = 0.79$). In another study, Dako (www.accessdata.fda.gov/cdrh_docs/pdf15/P150013c.pdf) evaluated the reproducibility of PD-L1 staining of their antibody 22C3 at 50% cut point. Sixty-two samples were assessed by three pathologists and the reproducibility was higher than in our study with an OPA of 92.7% (95% CI, 88.1–96.8) for the interobserver reproducibility and 96.4% (95% CI, 94.3–98.6) for the intraobserver reproducibility. Lastly, only one other study has assessed reproducibility of the PD-L1 staining with a number of pathologists similar to our study (29) using a variety of PD-L1 antibodies including the Dako 22C3 antibody used in our study. They assessed cut points for positivity ranging from 1% to 50% and Light's κ scores showed moderate concordance levels for all cut points: 0.74 (95% CI, 0.44–0.94) for the 1% cut point and 0.66 (95% CI, 0.42–0.89) for the 50% cut point with the 22C3 PD-L1 antibody. Similar concordance scores were found using other PD-L1 antibodies [E1L3N (Cell Signaling Technology), SP142 (Spring Bioscience Corporation), Dako 28-8, Ventana SP142, and Ventana SP263]. Our study reports results on the reproducibility of pathologists' assessments with a substantially higher number of both observers and samples, ensuring good reliability in terms of precision in the calculated values and robustness of the study results and are more likely to reflect real-life practice.

The training provided was found to have no or very little impact on the inter- or intraobserver reproducibility. A small impact on concordance between PPs' assessments and the gold standard was observed for samples with a very high PD-L1 tumor proportion

score (>80%). Education, training, and guidance have been reported to improve consistency in pathologist's assessments, but only up to a point beyond which improvement in methodology is needed (30). Our results suggest that the 1-hour training provided did not lead to significant improvement in pathologists' performance. PPs were representative of Australian national training and post fellowship experience in anatomical pathology reflecting the experience of most public and private laboratories. As PPs were experienced, with a median of 15 years of experience, PPs had been previously trained in assessing other immunohistochemistry biomarkers, such as ALK, HER2, and estrogen receptor. This could explain why the training had no impact on PPs' assessments. Training is not necessary for samples with a low PD-L1 tumor proportion score as agreement with the gold standard is already high (>80%). However, agreement with the gold standard was low for samples in the range 40%–60%. In the future, training pathologists to assess positive samples around the 50% cut point could help to improve the intra- and interobserver agreements for the assay and more training than can be provided in a 1-hour session could potentially be of benefit. This could be particularly important given a 50% threshold was used to select patients in the recently published clinical trial showing a survival advantage in patients treated first line with pembrolizumab compared with conventional chemotherapy (13). In particular, training should be conducted with a special focus on concomitant cytoplasmic staining or incomplete tumor membrane staining as both were reported by PPs as a cause of uncertainty during PD-L1 immunohistochemistry assessment. An external quality assessment scheme like the one that has been shown to be helpful for breast histopathology in the United Kingdom (30) could be developed including guidelines with examples, strategies for difficult cases, and on-line educational materials, to optimize pathologists' assessment of PD-L1 staining.

This study was designed to avoid potential bias identified at time of protocol development. Indeed, recall bias could have been observed as the PPs assessed the same samples twice over 2 days. However, recall bias was minimized because there were so many samples to be assessed and they were assessed in a different order for both evaluation sessions. Moreover, the sample sets were designed to avoid any bias influencing the statistical analysis by balancing PD-L1 positive and negative samples. Occurrence of statistical biases was assessed *a posteriori* using BAK and PABAK. Because no differences were observed for the values of Cohen's κ and the BAK, no index biases were observed. Despite all the precautions taken, the PABAK was not similar to the Cohen's κ coefficient for the 50% cut point, indicating possible prevalence bias. This bias may have arisen from the unbalanced distribution of positive and negative samples in the 50% cut point sample set. In addition, the gold standard as designed in this study was not entirely objective but consisted of subjective assessment of PD-L1 immunohistochemistry expression potentially weakening this study. However, the gold standard assessment was undertaken by highly trained pathologists, certified in scoring Dako 22C3 PD-L1, using the only method currently available that links tumor PD-L1 status with patient clinical response to anti-PD-1 treatment. Finally, using only tissue microarrays for evaluating the reproducibility of PD-L1 tumor proportion score assessment could be considered as the main limitation of our study. It could be useful to conduct additional studies using whole sections to confirm our results as this would be more representative of real clinical practice and adds complexities regarding which area to score.

In real clinical practice, samples to be assessed are not artificially enriched with samples close to cut points as was done in this study. According to previously published clinical trial data (20, 31), 61% of samples are positive considering a 1% cut point while only 23% would be strongly positive with a 50% cut point. In the gold standard, our study reported 24.3% of positive samples considering a 1% cut point and 10.3% of positive samples considering a 50% cut point. Thus it might be expected that most samples from patients with advanced stage NSCLC would have low PD-L1 positive staining, significantly increasing the OPA in the real clinical practice. Differences in the prevalence of PD-L1 positivity between the gold standard and previously published clinical data could result from the nature of the samples used in this study, including the use of tissue micro-arrays, early stage rather than late stage tumors, a high proportion of well differentiated tumors and potentially the age of the samples.

In our study, pathologists mostly underscored the samples in the 50% cut point sample set, which seemed to be due to a failure to assign cells with weak and/or incomplete membranous staining appropriately as positive. It is not unusual for the intensity of PD-L1 staining to be heterogeneous and its assessment is different from other recognized biomarkers (26). For example, weak/incomplete membranous straining is not included in scoring for HER2 staining, unlike for PD-L1 assessment. Thus, although the PPs were very experienced, they may have acquired their experience with scoring biomarkers other than PD-L1.

In conclusion, two different cut points were considered in this study, 1% and 50% as these are the cut points generally accepted with the 22C3 PD-L1 antibody (13, 20, 22). With a large number of assessed samples and a higher number of observers than in the previously published literature, our study demonstrated high intra- and interobserver reproducibility for the 1% cut point. Agreement was slightly lower for the 50% cut point, in particular for the interobserver variability, however, the Cohen's κ score (PABAK) was 0.64, indicating a substantial agreement. The lower reproducibility at the 50% cut point suggests that training and external quality assessment schema should focus on improving reproducibility of the samples with $\geq$50% of stained cells, particularly as this cut point will potentially be used in clinical practice to select NSCLC patients for anti-PD1 therapy.

Disclosure of Potential Conflicts of Interest

W.A. Cooper reports receiving commercial research support from Merck Sharp & Dohme, reports receiving speakers bureau honoraria from AstraZeneca, Bristol-Myers Squibb, and Merck Sharp & Dohme, and is a consultant/advisory board member for AstraZeneca, Bristol-Myers Squibb, and Merck Sharp & Dohme. P.A. Russell reports receiving compensation for participating in advisory boards for Bristol-Myers Squib and Merck. V. Sivasubramaniam reports receiving speakers bureau honoraria from Merck Sharp & Dohme and is a consultant/advisory board member for Bristol-Myers Squibb. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: W.A. Cooper, P.A. Russell, C. Faure, A. Reznichenko, A. Grattan, S.B. Fox

Development of methodology: W.A. Cooper, P.A. Russell, A. Reznichenko, S.B. Fox

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): W.A. Cooper, P.A. Russell, M. Cherian, E.E. Duhig, D. Godbolt, P.J. Jessup, C. Leslie, A. Mahar, D.F. Moffat, A. Reznichenko, A. Grattan, S.B. Fox

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): W.A. Cooper, P.A. Russell, C. Faure, S.B. Fox

Writing, review, and/or revision of the manuscript: W.A. Cooper, P.A. Russell, E.E. Duhig, P.J. Jessup, C. Leslie, D.F. Moffat, C. Faure, A. Reznichenko, A. Grattan, S.B. Fox
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): P.J. Jessup, V. Sivasubramaniam, A. Reznichenko, A. Grattan
Study supervision: W.A. Cooper, S.B. Fox.
Other (assessment of PD-L1 biomarker in NSCLC samples): C. Khoo

Acknowledgments

Authors (Wendy A. Cooper, Prudence A. Russell, Maya Cherian, Edwina E. Duhig, David Godbolt, Peter J. Jessup, Christine Khoo, Connull Leslie, Annabelle Mahar, David F. Moffat, Vanathi Sivasubramaniam, Celine Faure, Alena Reznichenko, Amanda Grattan, Stephen B. Fox) are responsible for the work described in this article. All authors were involved in at least one of the following: conception, design of work or acquisition, interpretation of data and drafting the manuscript and/or revising/reviewing the manuscript for important intellectual content. All authors provided final approval of the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Statistical support was provided by Philippe Huot-Marchand, MSc, of Mapi Group, Lyon, France and funded by MSD Australia.

Grant Support

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

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 17, 2017; revised February 13, 2017; accepted April 11, 2017; published OnlineFirst April 18, 2017.

References

1. International Agency for Research on Cancer. Globocan 2012: estimated cancer incidence, mortality and prevalence worldwide in 2012. Available from: http://globocan.iarc.fr/Default.aspx.
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65:5–29.
3. Scagliotti GV, Bironzo P, Vansteenkiste JF. Addressing the unmet need in lung cancer: the potential of immuno-oncology. Cancer Treat Rev 2015; 41:465–75.
4. Leighl NB. Treatment paradigms for patients with metastatic non-small-cell lung cancer: first-, second-, and third-line. Curr Oncol 2012;19(Suppl 1): S52–8.
5. Gerber DE, Schiller JH. Maintenance chemotherapy for advanced non-small-cell lung cancer: new life for an old idea. J Clin Oncol 2013;31:1009–20.
6. He J, Hu Y, Hu M, Li B. Development of PD-1/PD-L1 pathway in tumor immune microenvironment and treatment for non-small cell lung cancer. Sci Rep 2015;5:13110.
7. Teng MW, Ngiow SF, Ribas A, Smyth MJ. Classifying cancers based on T-cell infiltration and PD-L1. Cancer Res 2015;75:2139–45.
8. Konishi J, Yamazaki K, Azuma M, Kinoshita I, Dosaka-Akita H, Nishimura M. B7-H1 expression on non-small cell lung cancer cells and its relationship with tumor-infiltrating lymphocytes and their PD-1 expression. Clin Cancer Res 2004;10:5094–100.
9. Ilie M, Long-Mira E, Bence C, Butori C, Lassalle S, Bouhlel L, et al. Comparative study of the PD-L1 status between surgically resected specimens and matched biopsies of NSCLC patients reveal major discordances: a potential issue for anti-PD-L1 therapeutic strategies. Ann Oncol 2016;27:147–53.
10. Azuma K, Ota K, Kawahara A, Hattori S, Iwama E, Harada T, et al. Association of PD-L1 overexpression with activating EGFR mutations in surgically resected nonsmall-cell lung cancer. Ann Oncol 2014;25:1935–40.
11. Boland JM, Kwon ED, Harrington SM, Wampfler JA, Tang H, Yang P, et al. Tumor B7-H1 and B7-H3 expression in squamous cell carcinoma of the lung. Clin Lung Cancer 2013;14:157–63.
12. Velcheti V, Schalper KA, Carvajal DE, Anagnostou VK, Syrigos KN, Sznol M, et al. Programmed death ligand-1 expression in non-small cell lung cancer. Lab Invest 2014;94:107–16.
13. Reck M, Rodriguez-Abreu D, Robinson AG, Hui R, Csoszi T, Fulop A, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N Engl J Med 2016;375:1823–33.
14. Hino R, Kabashima K, Kato Y, Yagi H, Nakamura M, Honjo T, et al. Tumor cell expression of programmed cell death-1 ligand 1 is a prognostic factor for malignant melanoma. Cancer 2010;116:1757–66.
15. Wang SF, Fouquet S, Chapon M, Salmon H, Regnier F, Labroquere K, et al. Early T cell signalling is reversibly altered in PD-1+ T lymphocytes infiltrating human tumors. PLoS One 2011;6:e17621.
16. Dong H, Strome SE, Salomao DR, Tamura H, Hirano F, Flies DB, et al. Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion. Nat Med 2002;8:793–800.
17. Liu J, Hamrouni A, Wolowiec D, Coiteux V, Kuliczkowski K, Hetuin D, et al. Plasma cells from multiple myeloma patients express B7-H1 (PD-L1) and increase expression after stimulation with IFN-{gamma} and TLR ligands via a MyD88-, TRAF6-, and MEK-dependent pathway. Blood 2007;110: 296–304.
18. Patel SP, Kurzrock R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. Mol Cancer Ther 2015;14:847–56.
19. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627–39.
20. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. N Engl J Med 2015;372:2018–28.
21. Chen L, Han X. Anti-PD-1/PD-L1 therapy of human cancer: past, present, and future. J Clin Invest 2015;125:3384–91.
22. Langer CJ, Gadgeel SM, Borghaei H, Papadimitrakopoulou VA, Patnaik A, Powell SF, et al. Carboplatin and pemetrexed with or without pembrolizumab for advanced, non-squamous non-small-cell lung cancer: a randomised, phase 2 cohort of the open-label KEYNOTE-021 study. Lancet Oncol 2016;17:1497–508.
23. Gettinger S, Rizvi NA, Chow LQ, Borghaei H, Brahmer J, Ready N, et al. Nivolumab monotherapy for first-line treatment of advanced non-small-cell lung cancer. J Clin Oncol 2016;34:2980–7.
24. Fehrenbacher L, Spira A, Ballinger M, Kowanetz M, Vansteenkiste J, Mazieres J, 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.
25. Cooper WA, Tran T, Vilain RE, Madore J, Selinger CI, Kohonen-Corish M, et al. PD-L1 expression is a favorable prognostic factor in early stage non-small cell carcinoma. Lung Cancer 2015;89:181–8.
26. Harris LN, Ismaila N, McShane LM, Andre F, Collyar DE, Gonzalez-Angulo AM, et al. Use of biomarkers to guide decisions on adjuvant systemic therapy for women with early-stage invasive breast cancer: American Society of Clinical Oncology Clinical Practice Guideline. J Clin Oncol 2016;34:1134–50.
27. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977;33:363–74.
28. Phillips T, Simmons P, Inzunza HD, Cogswell J, Novotny JJr, Taylor C, et al. Development of an automated PD-L1 immunohistochemistry (IHC) assay for non-small cell lung cancer. Appl Immunohistochem Mol Morphol 2015;23:541–9.
29. Scheel AH, Dietel M, Heukamp LC, Johrens K, Kirchner T, Reu S, et al. Harmonized PD-L1 immunohistochemistry for pulmonary squamous-cell and adenocarcinomas. Mod Pathol 2016;29:1165–72.
30. Ellis IO, Coleman D, Wells C, Kodikara S, Paish EM, Moss S, et al. Impact of a national external quality assessment scheme for breast pathology in the UK. J Clin Pathol 2006;59:138–45.
31. Herbst RS, Baas P, Kim DW, Felip E, Perez-Gracia JL, Han JY, 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.