Accepted Manuscript

PD-L1 immunohistochemistry comparability study in real-life clinical samples: results of Blueprint phase 2 project

Ming Sound Tsao, MD, Keith M. Kerr, MD, Mark Kockx, MD, PhD, Mary-Beth Beasley, MD, Alain C. Borczuk, MD, Johan Botling, MD, Lukas Bubendorf, MD, Lucian Chirieac, MD, Gang Chen, MD, Teh-Ying Chou, MD, PhD, Jin-Haeng Chung, MD, PhD, Sanja Dacic, MD, PhD, Sylvie Lantuejoul, MD, Mari Mino-Kenudson, MD, Andre L. Moreira, MD, Andrew G. Nicholson, DM, Masayuki Noguchi, MD, PhD, Giuseppe Pelosi, MD, Claudia Poleri, MD, Prudence A. Russell, MD, Jennifer Sauter, MD, Erik Thunnissen, MD, PhD, Ignacio Wistuba, MD, PhD, Hui Yu, MD, PhD, Murry W. Wynes, PhD, Melania Pintilie, MSc, Yasushi Yatabe, MD, PhD, Fred R. Hirsch, MD, PhD.

PII: S1556-0864(18)30626-9
DOI: 10.1016/j.jtho.2018.05.013
Reference: JTHO 984

To appear in: Journal of Thoracic Oncology

Received Date: 29 April 2018
Revised Date: 12 May 2018
Accepted Date: 15 May 2018


This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
PD-L1 immunohistochemistry comparability study in real-life clinical samples: results of Blueprint phase 2 project

Ming Sound Tsao, MD,1 Keith M. Kerr, MD,2 Mark Kockx, MD,3 PhD, Mary-Beth Beasley MD,4 Alain C. Borczuk MD,5 Johan Botling MD,6 Lukas Bubendorf, MD,7 Lucian Chirieac, MD,8 Gang Chen, MD,9 Teh-Ying Chou MD, PhD,10 Jin-Haeng Chung MD, PhD,11 Sanja Dacic MD, PhD,12 Sylvie Lantuejoul MD,13 Mari Mino-Kenudson MD,14 Andre L. Moreira MD,15 Andrew G. Nicholson DM,16 Masayuki Noguchi MD, PhD,17 Giuseppe Pelosi MD,18 Claudia Poleri MD,19 Prudence A. Russell MD,20 Jennifer Sauter MD,21 Erik Thunnissen MD, PhD,22 Ignacio Wistuba MD, PhD,23 Hui Yu MD, PhD,24 Murry W. Wynes PhD,25 Melania Pintilie MSc,26 Yasushi Yatabe MD, PhD,27 Fred R. Hirsch MD, PhD.28

1Department of Pathology, University Health Network, Princess Margaret Cancer Centre and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario M5G 2C4, Canada

2Department of Pathology, Aberdeen Royal Infirmary, Aberdeen University Medical School, Aberdeen, Scotland, UK, AB25 2ZD

3HistoGeneX, Antwerp, Belgium

4Department of Pathology, Mount Sinai Medical Center, New York, NY, U.S.A.

5Department of Pathology, Weill Cornell Medicine, NY, New York, U.S.A.10065
6 Department of Immunology Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

7 Fachbereichsleiter Zytopathologie, Universitätsspital Basel, Pathologie, Basel, Switzerland

8 Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, U.S.A.

9 Department of Pathology, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China

10 Division of Molecular Pathology, Department of Pathology and Laboratory Medicine, Taipei Veterans General Hospital, Taipei, Taiwan

11 Department of Pathology and Respiratory Center, Seoul National University Bundang Hospital, Seongnam city, Gyeonggi-do 13620, Republic of Korea

12 Department of Pathology University of Pittsburgh, PUH C-608, 200 Lothrop Street, Pittsburgh, PA 15213, U.S.A.

13 Department of Biopathology, Centre Léon Bérard, Lyon, France Grenoble-Alpes University, INSERM U1209/CNRS 5309, Institute for Advanced Biosciences, Grenoble, France

14 Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, U.S.A.

15 New York University Langone Health, Department of Pathology, 560 First Avenue, TH4-15B, New York, N.Y., U.S.A. 10016
16 Department of Histopathology, Royal Brompton and Harefield NHS Foundation Trust and National Heart and Lung Institute, Imperial College, London, United Kingdom

17 Department of Pathology, Faculty of Medicine, University of Tsukuba, Tsukuba, Japan

18 Department of Oncology and Hemato-Oncology, University of Milan, and Inter-Hospital Pathology Division, IRCCS MultiMedica, Milan, Italy

19 Consultorio de Patología Especializada, Ciudad de Buenos Aires, Argentina

20 St Vincent's Pathology, Fitzroy, VIC 3065, Australia

21 Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, New York, U.S.A.

22 Department of Pathology, VU University Medical Center, Amsterdam, the Netherlands

23 Department of Translational Molecular Pathology, MD Anderson Cancer Center, Houston, TX, U.S.A. 77030

24 Division of Medical Oncology, University of Colorado Anschutz Medical Campus, Aurora, Colorado, U.S.A.

25 Department of Scientific Affairs, International Association for the Study of Lung Cancer, Aurora, CO, U.S.A.

26 Department of Biostatistics, University Health Network, Princess Margaret Cancer Centre Toronto, Ontario M5G 2C4, Canada

27 Department of Pathology and Molecular Diagnostics, Aichi Cancer Center, Nagoya, Japan
University of Colorado Cancer Center, International Association for the Study of Lung Cancer, Aurora, Colorado, U.S.A.

**Short heading:** Blueprint 2 PD-L1 IHC comparability study

**Keywords:** Immuno-oncology, checkpoint inhibitors, PD-1, companion diagnostics, complementary diagnostics, biomarker, assays, IASLC, lung cancer, NSCLC, cytology, pathology.

Address all correspondences to Dr. Fred R. Hirsch at Fred.Hirsch@ucdenver.edu
DISCLOSURES:

Ming Sound Tsao reports grants and personal fees from AstraZeneca, grants and personal fees from Merck, personal fees from BMS, personal fees from Ventana/Roche, grants and personal fees from Pfizer, outside the submitted work.

Keith M Kerr, MD reports personal fees from BMS, personal fees from MSD, personal fees from Merck Serono, personal fees from Roche, personal fees from AZ, outside the submitted work.

Mark Kockx, MD, PhD has nothing to disclose.

Mary-Beth Beasley, MD reports pathology consulting work from Genentech, Bristol Myers Squibb, and Merck, outside the submitted work.

Alain Borczuk, MD has nothing to disclose.

Johan Botling, MD reports personal fees from Astra Zeneca, personal fees from Merck Sharp & Dohme, personal fees from Roche, personal fees from Pfizer, personal fees from Bristol-Myers Squibb, personal fees from Boehringer Ingelheim, personal fees from Novartis, outside the submitted work.

Lukas Bubendorf, MD reports grants and personal fees from Roche, grants and personal fees from MSD, personal fees from BMS, personal fees from Astra zeneca, during the conduct of the study.

Lucian Chirieac, MD has nothing to disclose.

Gang Chen, MD, PhD has nothing to disclose.

Teh-Ying Chou, MD, PhD has nothing to disclose.

Jin Haeng Chung, MD, PhD has nothing to disclose.

Sanja Dacic, MD, PhDv reports personal fees from Bristol-Myers-Squibb, personal fees from Astra-Zeneca, outside the submitted work.

Sylvie Lantuejoul, MD, PhD reports grants and personal fees from Bristol Myers Squibb (BMS), personal fees from Merck Sharp and Dohme (MSD), personal fees from Roche, personal fees from Novartis, outside the submitted work.

Mari Mina-Kenudson, MD reports personal fees from Merrimack Pharmaceuticals, personal fees from H3 Biomedicine, personal fees from ACD, personal fees from Roche, outside the submitted work.

Andre L. Moreira, MD, PhD reports personal fees from Genetech, outside the submitted work.

Andrew G. Nicholson, DM reports personal fees from MERCK, personal fees from BOEHRINGER INGELHEIM, grants and personal fees from PFIZER, personal fees from NOVARTIS, personal fees from ASTRA ZENECA, personal fees from BRISTOL MYER SQUIBB, personal fees from ROCHE, personal fees from ASTRA ZENECA, personal fees from ABBVIE, outside the submitted work.

Masayuki Noguchi, MD, PhD has nothing to disclose.

Giuseppe Pelosi, MD has nothing to disclose.
Claudia Poleri, MD reports personal fees from ASTRA ZENECA, personal fees from Merck, personal fees from Roche, personal fees from BMS, outside the submitted work.

Prudence A. Russell, MD has nothing to disclose.

Jennifer Sauter, MD reports stock ownership in Merck, Thermo Fischer Scientific and Chemed Corporation, outside the submitted work.

Erik Thunnissen, MD, PhD reports personal fees from histogenex, personal fees from Ventana Roche, during the conduct of the study.

Ignacio Wistuba, MD, PhD reports grants and personal fees from Genentech/Roche, grants and personal fees from Bristol-Myers Squibb, personal fees from Boehringer Ingelheim, personal fees from Medscape, grants and personal fees from Astra Zeneca/Medimmune, grants and personal fees from Pfizer, grants and personal fees from HTG Molecular, personal fees from Asuragen, personal fees from GlaxoSmithKline, grants and personal fees from Merck, personal fees from Bayer, grants from Oncoplex, grants from DepArray, grants from Adaptive, grants from Adaptimmune, grants from Takeda, grants from Amgen, grants from EMD Serono, outside the submitted work.

Hui Yu, MD, PhD has nothing to disclose.

Murry W. Wynes, PhD has nothing to disclose.

Melania Pintilie, MSc has nothing to disclose.

Yasushi Yatabe, MD, PhD reports personal fees from MSD, personal fees from Chugai-pharm, personal fees from AstraZeneca, personal fees from Pfizer, personal fees from Novartis, outside the submitted work.

Fred R. Hirsch, MD, PhD is co-inventor of a University of Colorado owned patent: "EGFR IHC and- FISH as predictive biomarkers for EGFR Therapy". Dr. Hirsch has participated in advisory boards for: BMS, Genentech/Roche, HTG, Lilly, Merck, Pfizer, Ventana. Dr. Hirsch’s laboratory has received research grants (through University of Colorado) from: Genentech, BMS, Lilly, Bayer and Clovis.
Abstract

The Blueprint (BP) PD-L1 immunohistochemistry (IHC) comparability project is a pivotal academic/professional society and industrial collaboration to assess the feasibility of harmonizing the clinical use of five independently developed commercial PD-L1 IHC assays. The goals of BP phase 2 (BP2) were to validate the results obtained in BP1, using real world clinical lung cancer samples (n=81) of various histological and sample types, with all five trial-validated PD-L1 assays (22C3, 28-8, SP142, SP263, and 73-10), by an international panel of pathologists. BP2 also assessed the reliability of PD-L1 scoring by digital images, and on samples prepared for cytology examination. PD-L1 expression was assessed for percentage (tumor proportional scoring or TPS) tumor cells (TC) and immune cell (IC) areas showing PD-L1 staining, with TC scored continuously or categorically with cut-offs used in checkpoint inhibitor trials. The results showed highly comparable staining by 22C3, 28-8 and SP263 assays, less sensitivity with SP142 assay, and higher sensitivity with 73-10 assay to detect PD-L1 expression on TC. Glass slide and digital image scorings were highly concordant (Pearson correlation >0.96). There was very strong reliability among pathologists in TC PD-L1 scoring with all assays (overall intraclass correlation coefficient/ICC 0.86-0.93), poor reliability in IC PD-L1 scoring (overall ICC 0.18-0.19), and good agreement in assessing PD-L1 status on cytology cell blocks materials (ICC 0.78-0.85). BP2 consolidates the analytical evidence for interchangeability of 22C3, 28-8, and SP263 assays and lower sensitivity of SP142 assay for scoring TPS on TC, and demonstrates greater sensitivity of 73-10 compared to other assays.
Introduction

Immune checkpoint inhibitor therapies targeting the programmed death (PD)-1/PD-ligand (L)1 have become part of the standard of care in oncology. At least five inhibitors (nivolumab, pembrolizumab, atezolizumab, durvalumab and avelumab) have been approved by drug regulatory bodies in one or more countries for the treatment of several tumor types and for various indications. For advanced non-small cell lung cancer (NSCLC) patients without driver mutations (e.g. EGFR, ALK, ROS1 and BRAF) treatable by approved targeted therapies, nivolumab, pembrolizumab and atezolizumab are all available as second line treatment with (for pembrolizumab) or without (for nivolumab and atezolizumab) biomarker selection. Pembrolizumab is available for first-line monotherapy but only in patients with high PD-L1 expression, and in some countries, in combination with chemotherapy without any biomarker selection. Importantly, almost all clinical trials involving these inhibitors have demonstrated consistent correlation between their response rates and outcomes and the tumor cell PD-L1 expression levels, as measured by PD-L1 immunohistochemistry (IHC).

Therefore, despite only pembrolizumab requiring a PD-L1 IHC assay as a companion diagnostic to determine patient eligibility for treatment as approved by the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA), PD-L1 IHC has also been established as a complementary diagnostic for nivolumab and atezolizumab in non-squamous NSCLC and NSCLC, respectively.

As each PD-L1 IHC assay was independently developed for specific anti-PD-1/PD-L1 therapy using a different PD-L1 diagnostic assays (primary antibody clone plus
immunostaining platform/protocol), each assay potentially demonstrates distinct staining properties, which could prohibit the interchangeability of their clinical use. This would pose a significant challenge for pathology laboratories to offer PD-L1 testing, both from laboratory resources and budgetary points of view. Several groups have conducted studies to assess the comparability of the various PD-L1 IHC assays, and their potential interchangeability in clinical adoption, with all studies demonstrating similar results. The results from Blueprint phase 1 (BP1) study demonstrated that three PD-L1 assays (22C3, 28-8, SP263) showed comparable analytical performance for assessment of PD-L1 expression on tumor cells, while the SP-142 PD-L1 assay appeared to stain fewer tumor cells compared to the other assays. In contrast, all assays stained tumor infiltrating immune cells, but with poor concordance between assays. The BP-1 study had several limitations: (1) samples were obtained from a commercial source and did not necessarily reflect the real-world samples tested clinically, and (2) only three pathologists were involved in the scoring. Since BP-1, a fifth PD-L1 assay, using the 73-10 clone, has been developed as a potential assay for avelumab.

The goals of Blueprint phase 2 (BP2) are: (1) validate the assay comparability results obtained in BP1 using real world clinical lung cancer samples and all five trial-validated PD-L1 assays (22C3, 28-8, SP142, SP263, and 73-10) by a large international panel of clinically active pathologists, (2) assess the feasibility of PD-L1 scoring using digital images accessed by a web-based system, and (3) assess the reliability to score PD-L1 expression using samples prepared for cytology examination.

Materials and Methods:
**Materials**

Using respective institutional research ethics board approval, eighteen pathologists contributed eight unstained serial sections prepared from paraffin blocks of 81 lung cancer cases collected through their routine clinical practice. The final cases included 39 adenocarcinomas, 26 squamous cell carcinomas, 6 poorly differentiated non-small cell carcinomas, and 10 small cell carcinomas (Supplementary Table S1). The cases included 21 resections, 20 core needle or bronchial biopsies, 18 tumor positive lymph node excision biopsies or resections, and 22 cytology cell blocks.

The PD-L1 IHC 22C3 pharmDx assay and PD-L1 28-8 pharmDx assays were purchased from Dako (Heverlee, Belgium). VENTANA PD-L1 (SP142) Assay and VENTANA PD-L1 (SP263) Assay were purchased from Ventana (Tucson, AZ, USA).

**PD-L1 immunohistochemistry staining**

Each slide set of 81 cases were stained in a CLIA-approved IHC laboratory HistoGeneX (Antwerp, Belgium), using the FDA-approved 22C3, 28-8, SP142 and SP263 assays, and their respective protocols detailed in the product inserts and autostainers (Dako Autostainer Link48 for 22C3, 28-8 and 73-10, and Ventana BenchMark Ultra for SP142 and SP263). The PD-L1 73-10 assay was the protocol developed by Dako/Agilent (Santa Clara, CA, USA) for Avelumab clinical trials and transferred to the HistoGeneX. All immuno-stained slides and matching H&E stained sections were scanned by 3DHISTECH Pannoramic 250 Flash III at 20x magnification,
and the scanned images were uploaded and scored on the International Association for the Study of Lung Cancer (IASLC) server in Denver, CO, USA. Digital scoring was performed by accessing these images via the Pathomation Digital Pathology System (HistoGeneX).

**Scoring of PD-L1 assays**

The slides were scored by 24 experienced pulmonary pathologists (IASLC Pathology Committee members) from 15 countries across 5 continents. Since only some but not all participants have received company-sponsored assay-specific trainings, all participants were required to undergo 1.5 days pre-study group trainings by two experts from HistoGeneX for the scoring of PD-L1 IHC on tumor cells (TC) and immune cells (IC), as part of this project. As PD-L1 scoring on tumor cells are identical for all assays, training on TC scoring was not assay specific. Greater effort was devoted to training for IC scoring using SP-142 stained cases and the SP142 IC scoring algorithm. PD-L1 stained TC were scored as tumor proportional score (TPS), which represents the best estimated percentage (0% to 100%) of TC showing partial or complete membranous PD-L1 staining, and also into one of 7 categories (<1%, 1-4%, 5-9%, 10-24%, 25-49%, 50-79%, 80-100%). These categories represent cut-offs that have been used in various immune checkpoint inhibitor trials or suggested by the sponsors (e.g. ≥80% for avelumab). All assays were also scored for IC PD-L1 staining based on a pattern scoring method developed by HistoGeneX, adapted from the scoring approach described in the Ventana SP142 PD-L1 IHC assay brochure and detailed in Supplementary Figure S1. As only one set of glass slides was available for
each assay, each IASLC pathologist and one of the trainers were randomly assigned to conduct the scoring of two assays using the microscope (glass slide reading) and three assays by web-based digital images. The trainer’s scores were used as the standard reference score set.

Statistical analyses

Intra-class correlation coefficient (ICC) was used to assess scoring reliability for continuous TPS scores, and Fleiss Kappa statistics (FKS) for categorical scores after dichotomization based on specified cut-offs. ICC between 0.75-0.9 and >0.9 were considered as good and excellent reliability, respectively. Kappa scores of ≥0.9 were considered near perfect, 0.80-0.89 as strong, 0.70-0.79 as moderate, 0.40-0.69 as weak. The reliability of PD-L1 scoring was an assessment across all pathologists (excluding the trainer) to compare for each assay, and separately for digital and glass image scores. The mean TPS of all pathologists was used to assess the reliability of their scoring relative to the trainers score, and the agreement of digital vs. glass slide scoring used the Pearson correlation and graphical approach as described by Bland and Altman.

RESULTS

At the cut-off date for completion of scoring, 114 datasets were available for analyses. These include 50 datasets from glass slide (2 assays by 24 IASLC
pathologists and one trainer) and 74 datasets from digital image scoring (3 assays by 23 IASLC pathologists and 5 assays by the trainer).

Reliability of pathologists to score tumor cell PD-L1 expression

Overall the ICC among all pathologists for glass slide reading ranged from 0.88 to 0.93, and ICC for digital image reading ranged from 0.80 to 0.91, demonstrating very good to excellent reliability (Table 1). Comparable or slightly improved scores were obtained when only scores from NSCLC tissue were analyzed after excluding scores for small cell and cytology samples. High-level reliability (kappa >0.7) was also demonstrated using Kappa statistics at various cut-offs, especially those ≥ 5%, 10%, 25% and 50%, for both glass slide and digital readings (Figure 1 and supplementary Figure S2). Reliability was, however, slightly diminished at the 1% and 80% cut-offs, especially by digital image reading. We further demonstrated using the trainer’s scores as reference, that pathologists’ scores were strongly comparable (Supplementary Figure S3).

Comparability of tumor cell PD-L1 scoring by digital image vs glass slide readings

Due to logistic and timeline challenges, it was not possible to obtain matching digital and glass slide scores from each pathologist. Therefore, to assess the reliability of PD-L1 scoring by digital images compared to the standard microscopic assessment on glass slides, we used two statistical methods using the means of scores by both methods for each assay. Both Pearson correlation and Bolt and Altman’s methods
demonstrated very high correlation and agreement between the two methods of PD-L1 IHC reading (supplementary Figure S4 and S5). These results justify the pooling of all data regardless of scoring method in subsequent analyses.

**Comparability of PD-L1 staining between five assays on tumor cell staining**

The mean value of TPS scores across all readers (including trainer’s), using pooled digital and glass slide scores were derived for each assay and plotted across the samples (Figure 2A). Three assays (22C3, 28-8 and SP263) showed close approximation between their respective best-fit curves. In contrast, SP142 showed less sensitivity (lower TPS scores), while 73-10 showed greater sensitivity (higher TPS scores) to detect PD-L1 expression in the same samples. Comparable results were noted for both the whole sample cohort and the cohort that excluded small cell carcinoma and cytology samples. Figures 3 shows a representative case that demonstrates the similarities and differences at microscopy levels in staining intensity. Pair-wise comparison demonstrated the closest similarity between 22C3 and 28-8, and consistently greater sensitivity for 73-10 versus 22C3, 28-8 and SP263 (Figure 2B). The SP263 curve and the Bland and Altman plots do infer a slightly greater sensitivity in staining when compared to the 22C3 and 28-8 curves (Figure 2B and Supplemental Figures S6 and S7).

**Reliability of pathologists to score immune cell PD-L1 expression**
As IC scoring for PD-L1 could only be done on tissue sections, we focused the analyses only on NSCLC tissue samples only, excluding the cytology aspirate samples. Furthermore, as IC was scored categorical, Fleiss-Kappa (F-K) was used to assess inter-pathologist scoring reliability. The F-K among all pathologists for glass slide reading ranged from 0.11 to 0.28, and F-K for digital image reading ranged from 0.08 to 0.27, demonstrating overall poor agreement for assessment of PD-L1 staining on IC, using the system adopted in this study (Supplemental Table S2 and Figure S8). With both approaches, the highest overall reliability was achieved with SP-142 (F-K 0.27-0.28). There was weak to moderate agreement in scoring IC PD-L1 staining by using glass slide vs. digital scoring, and among pathologists vs. the trainer. Interestingly, moderate to strong agreement between pathologists vs. trainer was achieved for SP142 for distinguishing IC0 vs. IC1,2,3, but the agreements for higher IC categories were diminished (Supplemental Figure S8).

Comparability of PD-L1 staining between five assays on immune cell staining

Among the five assays, the distribution of IC scores among three assays (22C3, 28-8 and SP263) were comparable. In contrast, 73-10 and SP142 showed greater and lesser staining of the IC, respectively, compared to the other three assays (Supplemental Figure S9).

Reliability of PD-L1 scoring on tumor cells in cytology samples
Overall the ICC among all pathologists for reading cytology samples were good, both for glass slide (0.78) and digital (0.85) readings (Supplemental Table S3). However, these were slightly lower than those for NSCLC tissue-only samples (0.89 and 0.93, respectively) (Table 1). Comparable moderate levels of agreement (a majority with kappa >0.6) were noted at all cut-off levels (Figure 4), but overall these were lower than those achieved in non-cytology NSCLC tissue samples (Figure 1 and Supplemental Figure S2).

Discussion

This BP2 study using lung cancer diagnostic samples encountered in routine clinical pathology practice has further confirmed that three of the five currently available PD-L1 IHC assays (22C3 DAKO pharmDx, 28-8 DAKO pharmDx and Ventana SP263) show comparable staining characteristics on TC, while the Ventana SP142 assay shows less sensitivity and the DAKO pharmDx 73-10 IHC assay higher sensitivity to detect PD-L1 expression. We have also demonstrated that among a large group of pulmonary pathologists, the overall reliability or agreement in scoring PD-L1 was very strong especially on NSCLC tissue section samples (overall ICC >0.89). In contrast, we have shown that despite group training, scoring IC PD-L1 staining levels remains challenging with low ICC and poor kappa scores for all IC groups. Lastly, while the number of cases is limited, and the result is considered preliminary, we observed moderate agreement in the pathologists’ assessment of PD-L1 status in needle aspirate cell block specimens.
There have been several studies comparing the analytical performance of the commercially available PD-L1 IHC assays. While the designs, number and type of samples, and number of pathologists involved in these studies varied, practically all of them have reported high concordance in tumor cell staining between 22C3, 28-8 and SP263 assays. Almost all of the published studies including the BP1 have used resected NSCLC tissue. To our knowledge, BP2 is the first study that uses a mixture of samples that are routinely encountered in clinical practice including core needle/bronchial and lymph node biopsies, resection, and needle aspirate biopsy samples for cytology evaluation. We also included in our design small cell lung carcinoma (SCLC) samples, as at the time, the prevalence of PD-L1 expression in this type of tumor was unknown. However, among the 10 SCLC samples, only one showed staining for PD-L1, which is consistent with the reported PD-L1 expression in SCLC. It is important to note that none of the assays are currently approved to assess PD-L1 expression in SCLC. While this study was not designed to assess the distribution of PD-L1 expression across lung cancer cases, our samples included the full range of PD-L1 expression levels and reflect the frequency of PD-L1 negative cases, especially when lower stage, surgically resected cases are included. (Supplemental Figure S10).

BP1 results showed that the 22C3, 28-8 and SP263 assays had comparable sensitivity to detect PD-L1 expression on tumor cells, consistent with the results of three other studies, while SP142 showed significantly less sensitivity. Furthermore, in both the BP1 and a study by Ratcliffe, ≥90% agreements were achieved when staining of the same tumor sample was performed using SP263 versus 22C3 or 28-8 and the scoring was conducted by one pathologist. Perhaps because of these comparisons, the
SP263 PD-L1 IHC assay is CE-marked and available as a biomarker of nivolumab and pembrolizumab in Europe. However, our results detected slightly greater sensitivity for the SP263 assay compared to the 22C3 and 28-8 assays. It should be noted that both Scheel et al. and Hendry et al. also detected such a slightly greater sensitivity of SP263 compared to 22C3 and 28-8 assays. Whether this difference significantly affects the number of “PD-L1 positive” cases seen at lower cut-offs (e.g. 1%) or has an impact on clinical response rates, remains to be determined. (Supplemental Figure S11) In one study, the significant impact in terms of more patients over threshold was only seen for the 1% cut-off. However, Fujimoto et al. recently also reported good analytical concordance between 22C3, 28-8 and SP263 assays (weighted k coefficient 0.64-0.55) on 40 patient samples who have been treated by nivolumab, but they did not observe the higher staining sensitivity of SP263 compared to 22C3 or 28-8 assays, and the three assays demonstrated equivalent predictive performance of response to nivolumab using the receiving operating characteristics (ROC) analysis with area under the curve of 0.75-0.82.

BP2 study is the first publication that compares the staining characteristics of 73-10 assay to the other four existing PD-L1 assays and showed greater sensitivity for the 73-10 assay compared to all other assays. The implications of a significantly more sensitive assay are still to be clarified, but the transferability of such an assay, for general usage to select patients for agents other than avelumab must be in question, at least with the data currently available. In the EMR100070-005 trial, avelumab was used in first line NSCLC in a patient cohort selected by the 73-10 assay above the 80% cut off. Supplemental Figure S11 shows that, although the numbers are small, those
patients above this 80% cut off using 73-10, are well matched to the cohort selected by 22C3 above the 50% cut off. More data are undoubtedly required, but this raises the possibility of alternate biomarker/algorithim selection for avelumab.

In contrast to the other assays' scoring algorithm, the assessment of PD-L1 staining on IC in lung cancer is only included in the scoring algorithm of the SP142 PD-L1 IHC assay. The IC score is estimated as a percentage of tumor area (intratumoral and peri-tumoral desmoplastic stroma) with an IC infiltrate that shows PD-L1 staining, regardless of the type of IC (excepting macrophages in entrapped lung alveoli). Using this approach, several studies including the BP1 previously have reported greater variability and poor inter-pathologist concordance in the scoring of immune cell PD-L1 expression, with ICCs or kappa values of around 0.2. An alternate method (Supplemental Figure S1), still faithful to the concept of infiltrated tumor area but based on patterns of PD-L1 positive IC infiltrate has been proposed for scoring PD-L1 IC expression. As the SP142 assay often demonstrates a rather unique punctate type of staining not seen in other assays, we tested the feasibility of applying this approach to IC staining to the other assays in the BP2 project. The results demonstrate that reliability using NSCLC tissue sections was very poor (Fleiss Kappa values 0.11 to 0.21), despite extra emphasis during the group training on this scoring system and higher kappa values for SP142 (0.27-0.28). It is worth noting that the cases used in training included only large sections from resections samples, while this study included assessment of IC in metastatic lymph node and core/bronchial biopsy samples. This could have impacted on both the effectiveness and relevance of training received and on the reliability of IC scoring, as the assessment methods for IC are dependent on
spatial distribution of IC which is difficult to assess on small samples. The results emphasize the significant challenges in incorporating IC score in routine clinical testing. We also noted that IC staining in BP2 samples was slightly weaker with SP142 compared to 22C3, 28-8 and SP263, while staining with 73-10 was stronger. While this result was not seen in the BP1 study, it was noted in two other previous studies\textsuperscript{9,10}. If this observation is confirmed in other studies, the result strongly suggests a non-comparability of the other assays for assessing IC staining based on the SP142 scoring algorithm.

Two-thirds of non-small cell lung cancer patients present with advanced stage disease and a common diagnostic approach for these patients includes cytological assessment of fine needle aspiration biopsy or endobronchial ultrasound biopsy specimens. In some countries 50\% or more of such patients will have only cytology samples available for diagnosis.\textsuperscript{18} Unfortunately, none of the pivotal immune checkpoint inhibitor trials included cytology specimens in their biomarker program to develop the companion PD-L1 IHC assays. Thus, strictly speaking the current clinically used PD-L1 IHC assays has not been validated for all sample types encountered in their clinical application. Consequently, the diagnostic companies that commercialize the various PD-L1 assays have not recommended their use on cytology specimens, and the training offered to pathologists for these assays has not included examples of cytology specimens. While many pathologists consider no difference in their approach to assess PD-L1 expression on TC using cytology cell block or tissue biopsy/resection sections, only a few studies with small sample sizes have reported on the concordance of PD-L1 assessment on cytology vs. matching surgical specimens.\textsuperscript{19-22} In this first part of the
BP2 project, our results showed moderately good agreement (ICC 0.78-0.85 and kappa 0.6-0.85) among our group to score TC expression of PD-L1 cytology samples. Further confirmation of this is awaited in the next phase of BP2 that will compare the concordance of PD-L1 scores from fine needle aspirate vs. core biopsy vs. large section samples of the same lung resection specimen.

In conclusion, the BP2 results using real-life clinical lung cancer samples and scored by 25 pathologists have further affirmed the results of BP1, and also demonstrate that the new 73-10 assay being developed for avelumab shows greater sensitivity in detecting PD-L1 expression when compared with the other four assays. We have also confirmed the reliability of PD-L1 scoring by digital images, used by a large group of pulmonary pathologists in assessing PD-L1 expression on tumor cells but not on immune cells. Together with other published comparability studies, BP2 consolidates the evidence for interchangeability among 3 different assays (22C3, 28-8, and SP263), for use in tumor cell PD-L1 expression scoring (tumor proportion scoring or TPS), allowing a range of cut-offs matched with their respective therapeutic agents to be considered from the assessment of a single PD-L1 IHC test. Studies on PD-L1 staining heterogeneity and comparison of diagnostic values of large specimens versus small specimens versus cytology from the same tumors are ongoing in BP2B.

Acknowledgement

This study is supported by AstraZeneca, Bristol-Myers Squibb, Hoffmann La Roche/Genentech/Ventana Medical System, Merck & Co, Inc., EMD Serono/Merck KGaA/Pfizer, and Agilent Dako. We thank Kristine Brovsky and Chis Rivard in Dr. Fred
Hirsch’s lab, Dr. Wim Waelput and Dieter Rondas at Histogenex for providing crucial logistic supports. Dr. Pelosi wish to thank Dr. Silvano Bosari from the Division of Anatomic Pathology, Fondazione IRCCS Ca’ Granda Ospedale Maggiore Policlinico, Milan and Department of Pathophysiology and Transplantation, Università degli Studi di Milano, Milan, Italy, for assistance in providing four case materials for this study.
References


23

**Table 1.** Reliability (intraclass correlation coefficient) of scoring PD-L1 expression on tumor cells among all pathologists (excluding the trainer) for all cases and non-small cell lung cancer (NSCLC) biopsy/resected cases.

<table>
<thead>
<tr>
<th></th>
<th>Glass slide scoring</th>
<th>Digital scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All cases</td>
<td>NSCLC tissue only</td>
</tr>
<tr>
<td><strong>22C3</strong></td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td><strong>28-8</strong></td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td><strong>SP-142</strong></td>
<td>0.88</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>SP-263</strong></td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td><strong>73-10</strong></td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td><strong>All assays</strong></td>
<td>0.86</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Figure legends

**Figure 1.** Reliability of scoring tumor cells PD-L1 expression using Fleiss Kappa statistics at 1%, 25% and 50% cut-offs, for digital and glass slide readings, respectively, and for all cases (whole cohort) or NSCLC only with cytology excluded.

**Figure 2.** Comparability of PD-L1 staining on tumor cells (TC) among five assays. (A) overall comparison, and (B) pair-wise comparisons.

**Figure 3.** A representative case comparing the PD-L1 staining by five assays.

**Figure 4.** Agreement of scoring tumor cells PD-L1 expression in non-small cell lung cancer (NSCLC) cytology specimens at different cut-offs.
Figure 1.
Figure 2.

A. Whole cohort

B. NSCLC, Cytology excluded

Legend:
- 22C3
- 28–8
- SP142
- SP263
- 73–10
Figure 3.
Figure 4.