Modern Pathology (2020) 33:1746–1752
https://doi.org/10.1038/s41379-020-0544-x

ARTICLE

Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer

Emily S. Reisenbichler[1] · Gang Han[2] · Andrew Bellizzi[3] · Veerle Bossuyt[4] · Jane Brock[5] · Kimberly Cole[1] · Oluwole Fadare[6] · Omar Hameed[7,8] · Krisztina Hanley[9] · Beth T. Harrison[5] · M. Gabriela Kuba[10] · Amy Ly[4] · Dylan Miller[11,12] · Mirna Podoll[8] · Anja C. Roden[13] · Kamaljeet Singh[14] · Mary Ann Sanders[15] · Shi Wei[16] · Hannah Wen[9] · Vasiliki Pelekanou[1,17] · Vesal Yaghoobi[1] · Fahad Ahmed[1] · Lajos Pusztai[1] · David L. Rimm[1]

Received: 2 February 2020 / Revised: 25 March 2020 / Accepted: 25 March 2020 / Published online: 16 April 2020
© The Author(s), under exclusive licence to United States & Canadian Academy of Pathology 2020

Abstract

The US Food and Drug Administration (FDA) approved the PD-L1 immunohistochemical assay, SP142, as a companion test to determine eligibility for atezolizumab therapy in patients with advanced triple negative breast cancer (TNBC) but data in lung cancer studies suggest the assay suffers from poor reproducibility. We sought to evaluate reproducibility and concordance in PD-L1 scoring across multiple pathologists. Full TNBC sections were stained with SP142 and SP263 assays and interpreted for percentage (%) immune cell (IC) staining by 19 pathologists from 14 academic institutions. Proportion of PD-L1 positive cases (defined as ≥1% IC) was determined for each assay as well as concordance across observers. We utilized a new method we call Observers Needed to Evaluate Subjective Tests (ONEST) to determine the minimum number of evaluators needed to estimate concordance between large numbers of readers, as occurs in the real-world setting. PD-L1 was interpreted as positive with the SP142 assay in an average 58% of cases compared with 78% with SP263 ($p < 0.0001$). IC positive continuous scores ranged from 1 to 95% (mean = 20%) and 1 to 90% (mean = 10%) for SP263 and SP142, respectively. With SP142, 26 cases (38%) showed complete two category (<1% vs. ≥1%) concordance; with SP263, 38 cases (50%) showed complete agreement. The intraclass correlation coefficient (ICC) for two category scoring of SP263 and SP142 was 0.513 and 0.560. ONEST plots showed decreasing overall percent agreement (OPA) as observer number increased, reaching a low plateau of 0.46 at ten observers for SP263 and 0.41 at eight observers for SP142. IC scoring with both assays showed poor reproducibility across multiple pathologists with ONEST analysis suggesting more than half of pathologists will disagree about IC scores. This could lead to many patients either receiving atezolizumab when they are unlikely to benefit, or not receiving atezolizumab when they may benefit.

Supplementary information The online version of this article (https://doi.org/10.1038/s41379-020-0544-x) contains supplementary material, which is available to authorized users.

✉ David L. Rimm
David.Rimm@Yale.edu

1 Yale School of Medicine, New Haven, CT, USA
2 Texas A&M University, College Station, TX, USA
3 University of Iowa, Iowa City, IA, USA
4 Massachusetts General Hospital, Boston, MA, USA
5 Brigham and Women's Hospital, Boston, MA, USA
6 University of California San Diego, San Diego, CA, USA
7 Forward Pathology Solutions, Kansas City, MO, USA
8 Vanderbilt University Medical Center, Nashville, TN, USA
9 Emory University, Atlanta, GA, USA
10 Memorial Sloan Kettering Cancer Center, New York, NY, USA
11 Intermountain Healthcare, Salt Lake City, UT, USA
12 University of Utah, Salt Lake City, UT, USA
13 Mayo Clinic, Rochester, MN, USA
14 Brown University, Providence, RI, USA
15 University of Louisville, Louisville, KY, USA
16 University of Alabama at Birmingham, Birmingham, AL, USA
17 Sanofi Oncology US, Cambridge, MA, USA

Introduction

Invasive carcinoma that is negative for the expression of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2, also known as triple negative breast cancer (TNBC), is an aggressive form of breast cancer with few specific therapeutic targets. Recently, the IMpassion130 study demonstrated prolonged overall survival when atezolizumab, a PD-L1 inhibitor, was added to nab-paclitaxel in PD-L1 positive patients with advanced TNBC [1]. In the study, ~40% of tumors were PD-L1 positive, utilizing a cutoff of ≥1% immune cell (IC) staining with the Ventana SP142 immunohistochemical (IHC) assay. The U.S. Food and Drug Administration (FDA) approved this assay as a companion test to determine patient eligibility for atezolizumab therapy either on primary or metastatic tumor tissues. The FDA summary of safety and effectiveness data (SSED) for SP142 indicated high inter-laboratory reproducibility with nearly 95% overall percent agreement (OPA) between two readers for two category scoring of IC (positive vs. negative) in TNBC in a central laboratory [2]. However, literature in lung cancer, including a broader range of evaluators, shows that pathologists have low rates of agreement in assessing PD-L1 on IC [3–5]. PD-L1 testing for breast cancer is now becoming widespread in pathology laboratories with thousands of pathologists interpreting this stain. The goal of this study was to assess concordance in PD-L1 scoring between multiple pathologists from several different institutions with the SP142 and SP263 antibodies with no other training than following the manufacturer's instructions for scoring, as it is occurring currently in general pathology practice in the USA. We also present a new method to determine the minimum number of evaluators needed to obtain a representative estimate of concordance between large numbers of readers as occurs in routine clinical practice settings.

Materials and methods

Patient cohort and chromogenic immunohistochemistry

Slides representing primary invasive TNBCs (stage I-III) from 100 patients were obtained from the Yale School of Medicine Department of Pathology archives (Table 1) by selection of cases of TNBC accessioned between 2012 and 2016. Cases were reviewed and selected if they had sufficient tumor present to sustain 30–50 sections. For this set, 50 cases were from African American patients that were approximately matched by diagnosis date to 50 non-African American patients for another study (R01-CA219647). All tissues and data were retrieved under permission from the Yale Human Investigation Committee protocol #9505008219 to DLR. Full tissue sections from each case were stained with Ventana SP263 and SP142 commercial assays exactly according to manufacturer's instructions on the package insert using the Ventana Benchmark autostainer. Cell line arrays were used as positive control on separate slides from study cases. Following IHC staining, cases with insufficient invasive tumor cells on the tissue section or tissue folding were deemed uninterpretable and excluded from review.

Table 1 Patient characteristics at time of diagnosis.

	SP142 $n = 68$	SP263 $n = 76$
Patient age–med (range) Year	57 (33–84)	56 (32–90)
Race – No. (%)		
White	37 (54.4)	38 (50)
Black	31 (45.6)	38 (50)
Stained Tissue – No. (%)		
Primary tumor	68 (100)	76 (100)
Metastatic tumor	0 (0)	0 (0)
Pathologic Stage – No. (%)		
T1a	0 (0)	0 (0)
T1b	2 (2.9)	2 (2.6)
T1c	29 (42.6)	31 (40.8)
T2	33 (48.5)	39 (51.3)
T3	2 (2.9)	2 (2.6)
T4	2 (2.9)	2 (2.6)
NX	3 (4.4)	3 (3.9)
N0	41 (60.3)	44 (57.9)
N1	18 (26.4)	23 (30.2)
N2	5 (7.3)	5 (6.7)
N3	1 (1.5)	1 (1.3)

Slide scanning and pathologist scoring

Whole slide scanning of stained slides was performed at 20x using the Leica Aperio ScanScope, Controller v10.2.0.2359 and ScanScope Console v10.2.0.2352 imaging software. Digital files of the scanned images were distributed to 19 randomly selected peer pathologists from 14 institutions with a Yale constructed power point tutorial including the manufacturer's guidelines for the scoring system as well as representative photos available from the online Ventana product "Interpretation Guide" (https://productlibrary.ventana.com/ventana_portal/OpenOverlayServlet?launchIndex=1&objectId=740-48591018231EN). Pathologists were instructed to score both assays as the % IC staining over the tumoral area, as described in the package insert for the SP142 assay. Pathologists independently scored cases as either negative (<1% IC staining) or positive (≥1% IC staining) and estimated the percentage of IC staining for positive cases. Most pathologists had a primary or secondary

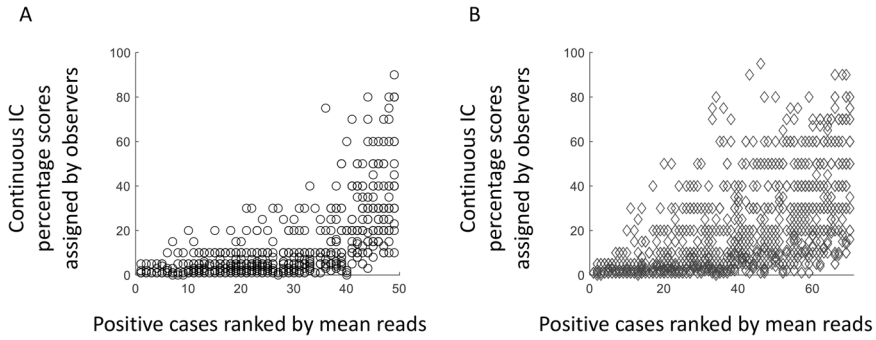

Fig. 1 Distribution of assigned continuous percentage scores. Percentage scores assigned by observers for positive cases (≥1% IC) with SP142 (**a**; $n = 49$ positive cases) and SP263 (**b**; $n = 70$ positive cases).

interest in breast pathology and participate in sign out of this subspecialty with 5–10 years of practice experience (see supplementary table 1). The few nonbreast pathologists were members of the College of American Pathologists Immunohistochemistry Committee.

Statistical analysis and observers needed to evaluate a subjective test (ONEST)

We invented a method that we call Observers Needed to Evaluate Subjective Tests (ONEST) to visualize the change in OPA as a function of the number of observers. For any combination of pathologists, we quantify the OPA using the proportion of tissue samples upon which all selected pathologists agree. Calculation of OPA for all permutations of 19 pathologists results in 19! (19 factorial) $= 1.22 \times 10^{17}$ combinations. We randomly select 100 permutations and plot the OPA against the number of pathologists. The resulting graphs descend to a plateau that begins at the number of pathologists we believe are required to provide realistic concordance estimates when the assay is broadly used. If the test is easy to interpret, resulting in high concordance among the observers, then the plateau will occur at a high OPA with a small number of observers (i.e., OPA estimates do not significantly change despite including more readers). In contrast, when there is high discordance amongst observers, then the plateau begins at a higher number of observers and it occurs at a lower OPA (see supplementary fig. 1). We believe that this approach could be used to evaluate any sort of subjective assessment, not just those seen in anatomic pathology.

Results

Following the exclusion of uninterpretable cases and cases excluded due to technical error, 68 cases were evaluable for SP142 and 76 for SP263 chromogenic staining (Table 1). PD-L1 was interpreted as positive with SP263 in an average of 78% of cases (range 64–96%) compared with 58% (range 41–81%) with the SP142 assay ($p < 0.0001$). Continuous scores for IC positive cases ranged from 1–95% (mean = 20%) with SP263 and 1–90% (mean = 10%) for SP142 (Fig. 1). Figure 1 illustrates the overall lower scoring seen with SP142 compared with SP263, consistent with previous reports in lung cancer [4, 5]. Continuous scores for SP142 showed less variability than the wide ranges of scores seen with SP263. The case with the largest variation included IC continuous scores from 10 to 90%. Complete two-category (<1% vs. ≥1%, i.e., negative vs. positive) scoring agreement across all observers was achieved in only 26 cases (38%) with SP142 and in 38 cases (50%) with SP263. A subset of cases showed a near even divide between being designated as positive or negative (Fig. 2a, c, arrows). Seven cases (11%) with SP142 and six cases (10%) with SP263, and were designated as negative by 40–60% of pathologists and positive by the remaining pathologists. On rare occasion, one of these cases was interpreted as showing >50% IC staining by some readers while <1% by others, but mostly, these cases showed levels of staining in single digits across readers. In these instances, some observers designated the staining as <1% and others between 1–10% leading to discordant positivity status assignment. Interestingly, the six cases with SP263 resulting in split interpretation by pathologists were entirely different from the seven cases with split results using the SP142 assay, indicating substantially different assay characteristics. Individual pathologists interpreted 41–81% of all cases as positive with SP142 and 64–96% of cases as positive with SP263 (Fig. 2b, d). The intraclass correlation coefficients (ICC) for two-category (<1% or ≥1%) scoring of SP142 and SP263 were 0.560 and 0.513, respectively. Continuous IC scores were used to further categorize cases into three-category (<1%; ≥1% but <10%; ≥10%) and four-category (<1%, ≥1% but <10%, ≥10% but <50%; ≥50%) scoring. The ICC remained similar for three-category (0.652 and 0.565) and four-category scoring (0.649 and 0.534) with SP142 and SP263 respectively.

ONEST plots for each assay for 19 pathologists with a two-category cut-point showed a decrease in OPA as the number of observers increased, reaching a plateau of ~0.46 at ten observers for SP263 and 0.41 at nine observers for SP142 (Fig. 3). As expected, three- and four-category

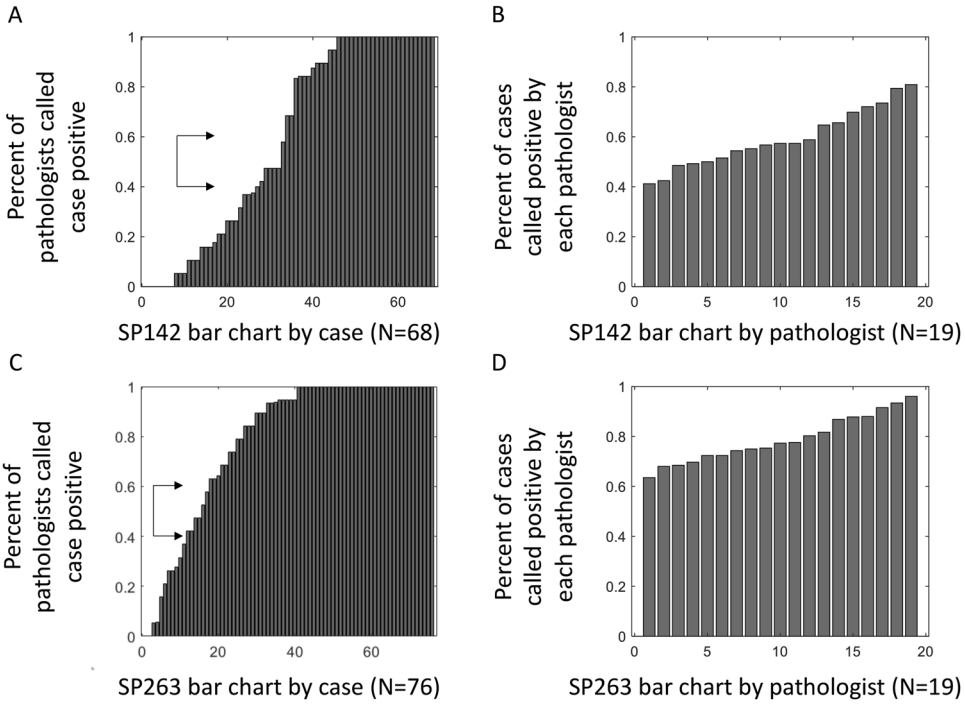

Fig. 2 Percentage of cases designated as positive by pathologists. Plot of the percentage of pathologists designating each individual case as positive (**a**) and the percentage of cases called positive by each of the 19 pathologists (**b**) when assessed with SP142. Frames (**c**) and (**d**) show the same distributions for the SP263 assay. Arrows in frames (**a**) and (**c**) designate those cases with near even split between positive and negative interpretation.

classifications showed much lower plateaus of agreement, quickly dropping below 20% OPA with more than two observers. With three or four categories, most comparisons involving more than six pathologists have zero OPA (i.e., no set of any six pathologists assigns the exact same category assignment to any single case).

Discussion

Utilizing the FDA approved SP142 assay and the recommended cutoff of ≥1% IC staining, we found a higher prevalence, average 58%, of PD-L1 positive tumors than published in the IMpassion130 trial (about 40%). Our finding is closer to the 49.7% prevalence published in the FDA SSED for SP142 [2]. The difference in prevalence between our study and the IMpassion130 trial is likely due to differences in tissue sources utilized for staining. The tissues utilized for PD-L1 staining in the published trial included both primary (~60%) and metastatic tumors (~40%), compared with our study which included only primary invasive tumor samples. Previous studies have shown that the number of TILs is decreased in distant metastases of breast cancer compared with primary tumors [6], and PD-L1 expression is lower in distant metastatic organ sites [6–9].

Although a cutoff for positive PD-L1 determination in breast cancer has not been established for SP263, utilizing the same 1% IC cutoff that is recommended with SP142, a 20% higher prevalence of positive cases was identified with SP263. This finding is in keeping with prior data showing SP142 to be less sensitive to detect the PD-L1 protein than other assays when evaluated in non-small cell lung carcinomas [4, 5] and by analysis of cell lines [10, 11]. The dissimilar proportion of positive cases seen with two different assays shows that these assays cannot be used interchangeably, highlighting the need to establish specific cutoffs for each assay that corresponds to clinical benefit from immunotherapy [12]. Recent post-hoc analysis of the IMpassion130 study cases found similar prevalence rates of PD-L1 positive cases with SP142 and SP263 assays [13].

Despite the inter-laboratory reproducibility studies shared by the FDA in the SSED showing overall agreement (compared with a consensus score) of >95%, we found substantially lower reproducibility across multiple pathologists with both assays. This illustrates a weakness of only using two observers or comparing a single observer to a consensus standard. To graphically illustrate this problem, we developed the ONEST plot. The ONEST analysis found a plateau of observer agreement at ~10 pathologists, reaching a stable OPA of around 0.4 for IC assessment for these two assays. This suggests that to better estimate the real word performance of assays, in terms of interobserver concordance, investigators and regulatory agencies need to use a larger number of observers. Our results also suggest that across many cases, more than half of the pathologists will assign discordant PD-L1 category to the same case. In a mutually exclusive two-category assignment (i.e., positive vs. negative), this implies that many cases could be assigned to the wrong category depending on the reader. This could

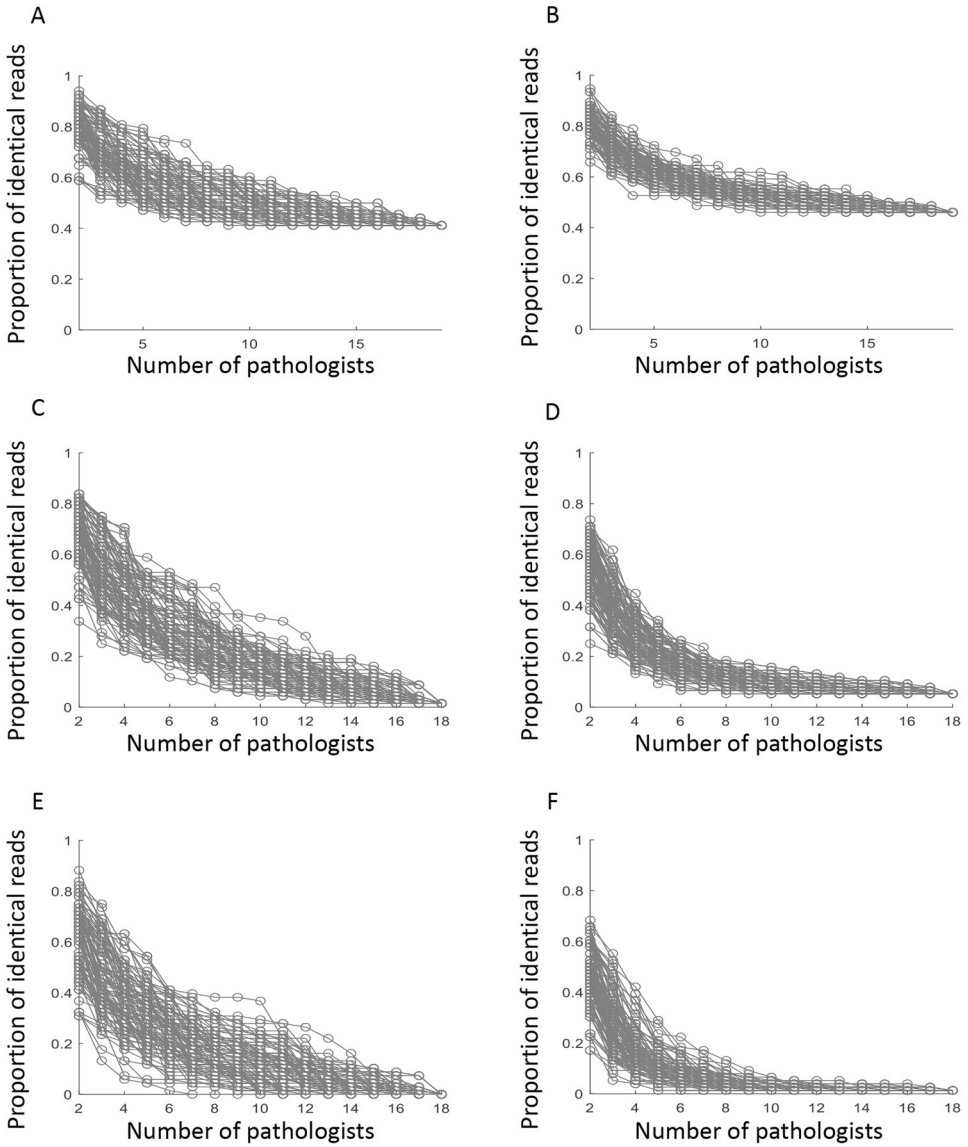

Fig. 3 ONEST plots showing overall percent agreement (OPA) or proportion of identical reads between pathologists as a function of the number of observers. One hundred curves were randomly selected from all possible combinations, using two category cutoffs for SP142 (**a**) and SP263 (**b**); three category cutoffs for SP142 (**c**) and SP263 (**d**); four category cutoffs for SP142 (**e**) and SP263 (**f**).

lead to a high percentage of patients either receiving atezolizumab, when they are unlikely to benefit, or not receiving atezolizumab when they may benefit.

Further review of specific cases resulting in the greatest range of IC scores revealed two possible sources of differences in interpretation. In some cases, there is unquestionable staining present but whether this represents staining of the TCs or intermingled ICs is less clear. In addition, tissue folding in areas of staining may also have contributed to problems distinguishing IC staining from background, edge artifact, or TC staining. This tissue folding may also increase difficulty in quantifying the area of staining. Each of these possible sources of discordance were seen with both assays as demonstrated in Fig. 4. The possibility of the review of whole slides images rather than glass slides leading to increased discordance was considered. However, the Blueprint Phase 2 project demonstrated that PD-L1 interpretation in non-small cell lung carcinoma results in high correlation and agreement between digital images and conventional glass slides [5]. FDA approval of whole slide imaging for primary diagnosis was based largely on studies demonstrating the noninferiority of whole slide imaging to glass slides in diagnostic concordance [14]. There is therefore no indication that glass slide review would improve concordance.

The high discordance rate in assessing IC PD-L1 positivity by either assay, which is consistent with previous studies in lung cancer, raises doubt about the ability of the unaided human eye to accurately and reproducibly quantify this feature. But, PD-L1 IC staining was strongly and significantly associated with benefit from Atezolizumab in the IMpassion130 trial. We suggest that this was due to central scoring by a small number of highly trained individuals. We believe other methods, perhaps automated assessment, or

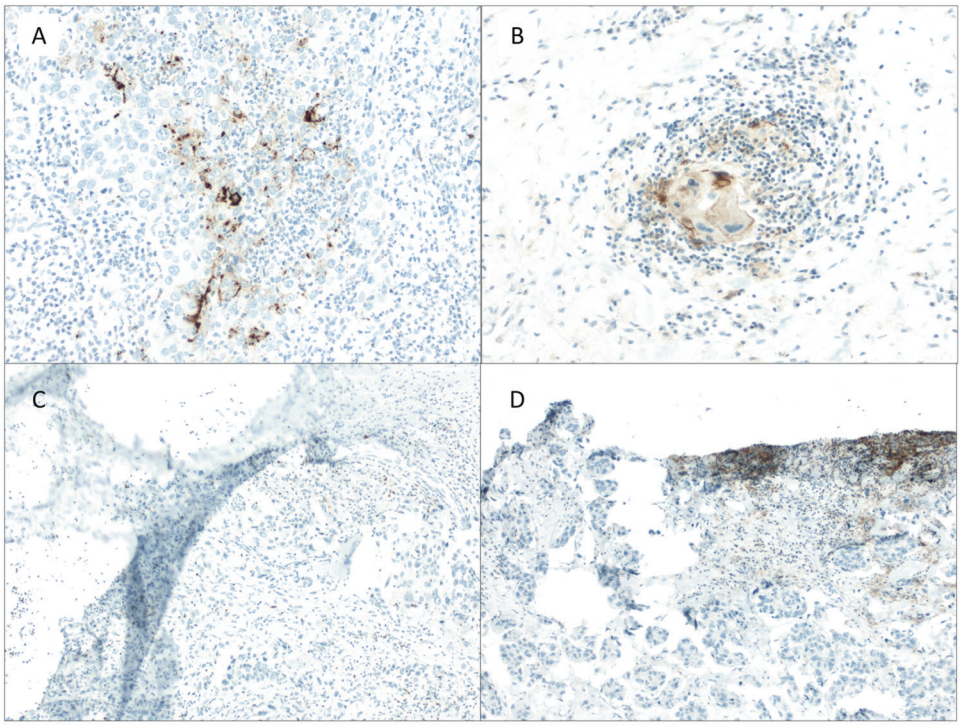

Fig. 4 Example cases with high discordance in immune cell scoring across observers. Staining of immune cells and admixed tumor cells with SP142 (**a**) and SP263 (**b**) (200×). Rolling of the tissue edge contributing to difficulty in assessment of staining with SP142 (**c**) and SP263 (**d**) (100×).

focused pathologist training and certification, need to be employed to make the assay successful in the clinical setting.

In summary, similar to what has been shown previously in non-small cell lung carcinomas, scoring of PD-L1 expression on IC is inconsistent across a large number of pathologists. Our ONEST method of analysis suggests that prior to test approval, an assay similar to this one should be interpreted by at least 8–10 pathologists rather than 2 or 3 to truly estimate assay reproducibility in the real-world setting. While PD-L1 testing currently plays a significant role in the management of an increasing number of advanced carcinomas, standardization of this subjective test has not been achieved and its current use in the clinic may result in patient harm due to misclassification of patients as PD-L1 negative who could benefit from therapy and exposing patients with false positive results to costly and potentially toxic therapy.

Acknowledgements This research was supported by the Breast Cancer Research Foundation (DLR and LP) and an NCI R01 grant (R01CA219647) to LP.

Compliance with ethical standards

Conflict of interest LP has received consulting fees and honoraria from Astra Zeneca, Merck, Novartis, Genentech, Eisai, Pieris, Immunomedics, Seattle Genetics, Almac and Syndax. DLR has served as an advisor for Astra Zeneca, Agendia, Amgen, BMS, Cell Signaling Technology, Cepheid, Daiichi Sankyo, Genoptix/Novartis, GSK, Konica Minolta, Merck, NanoString, PAIGE.AI, Perkin Elmer, Roche, Sanofi, Ventana and Ultivue. Astra Zeneca, Cepheid, NavigateBP, NextCure, Nanostring, Lilly, and Ultivue fund research in DLRs lab. VP is now an employee of Sanofi Aventis.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Schmid P, Adams S, Rugo HS, Schneeweiss A, Barrios CH, Iwata H, et al. Atezolizumab and nab-paclitaxel in advanced triple-negative breast cancer. N Engl J Med. 2018;379:2108–21.
2. Administration USFaD. Summary of Safety and Effectiveness Data (SSED) PMA P160002/S009. 2019. https://www.accessdata.fda.gov/cdrh_docs/pdf16/P160006B.pdf.
3. Hirsch FR, McElhinny A, Stanforth D, Ranger-Moore J, Jansson M, Kulangara K, et al. PD-L1 immunohistochemistry assays for lung cancer: results from phase 1 of the blueprint PD-L1 IHC assay comparison project. J Thorac Oncol. 2017;12:208–22.
4. Rimm DL, Han G, Taube JM, Yi ES, Bridge JA, Flieder DB, 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.
5. Tsao MS, Kerr KM, Kockx M, Beasley MB, Borczuk AC, Botling J, 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.
6. Ogiya R, Niikura N, Kumaki N, Yasojima H, Iwasa T, Kanbayashi C, et al. Comparison of immune microenvironments between primary tumors and brain metastases in patients with breast cancer. Oncotarget 2017;8:103671–81.
7. Cimino-Mathews A, Thompson E, Taube JM, Ye X, Lu Y, Meeker A, et al. PD-L1 (B7-H1) expression and the immune tumor microenvironment in primary and metastatic breast carcinomas. Hum Pathol. 2016;47:52–63.

8. Szekely B, Bossuyt V, Li X, Wali VB, Patwardhan GA, Frederick C, et al. Immunological differences between primary and metastatic breast cancer. Ann Oncol. 2018;29:2232–9.
9. Li Y, Chang C-W, Tran D, Denker M, Hegde P, Molinero L. Prevalence of PDL1 and tumor infiltrating lymphocytes (TILs) in primary and metastatic TNBC [abstract]. Cancer Res 2018;78 (4Suppl):Abstract nr PD6-01.
10. Martinez-Morilla S, McGuire J, Gaule P, Moore L, Acs B, Cougot D, et al. Quantitative assessment of PD-L1 as an analyte in immunohistochemistry diagnostic assays using a standardized cell line tissue microarray. Lab Investig. 2020;100:4–15.
11. Toki MI, Mani N, Smithy JW, Liu Y, Altan M, Wasserman B, et al. Immune marker profiling and programmed death ligand 1 expression across NSCLC mutations. J Thorac Oncol. 2018;13:1884–96.
12. Torlakovic E, Lim HJ, Adam J, Barnes P, Bigras G, Chan AWH, et al. "Interchangeability" of PD-L1 immunohistochemistry assays: a meta-analysis of diagnostic accuracy. Mod Pathol. 2020;33:4–17.
13. Rugo HS, Loi S, Adams S, Schmid P, Schneeweiss A, Barrios CH, et al. Performance of PD-L1 immunohistochemistry (IHC) assays in unresectable locally advanced or metastatic triple-negative breast cancer (mTNBC): post-hoc analysis of IMpassion130. Ann Oncol. 2019;30(suppl 5):v858–9.
14. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: a multicenter blinded randomized noninferiority study of 1992 cases (pivotal study). Am J Surg Pathol. 2018;42:39–52.