JAMA Oncology | Original Investigation

A Prospective, Multi-institutional, Pathologist-Based Assessment of 4 Immunohistochemistry Assays for PD-L1 Expression in Non–Small Cell Lung Cancer

David L. Rimm, MD, PhD; Gang Han, PhD; Janis M. Taube, MD; Eunhee S. Yi, MD; Julia A. Bridge, MD; Douglas B. Flieder, MD; Robert Homer, MD, PhD; William W. West, MD; Hong Wu, MD; Anja C. Roden, MD; Junya Fujimoto, MD; Hui Yu, MD; Robert Anders, MD; Ashley Kowalewski, MS; Christopher Rivard, PhD; Jamaal Rehman, MD; Cory Batenchuk, PhD; Virginia Burns, PhD; Fred R. Hirsch, MD, PhD; Ignacio I. Wistuba, MD, PhD

IMPORTANCE Four assays registered with the US Food and Drug Administration (FDA) detect programmed cell death ligand 1 (PD-L1) to enrich for patient response to anti–programmed cell death 1 and anti–PD-L1 therapies. The tests use 4 separate PD-L1 antibodies on 2 separate staining platforms and have their own scoring systems, which raises questions about their similarity and the potential interchangeability of the tests.

OBJECTIVE To compare the performance of 4 PD-L1 platforms, including 2 FDA-cleared assays, 1 test for investigational use only, and 1 laboratory-developed test.

DESIGN, SETTING, AND PARTICIPANTS Four serial histologic sections from 90 archival non–small cell lung cancers from January 1, 2008, to December 31, 2010, were distributed to 3 sites that performed the following immunohistochemical assays: 28-8 antibody on the Dako Link 48 platform, 22c3 antibody on the Dako Link 48 platform, SP142 antibody on the Ventana Benchmark platform, and E1L3N antibody on the Leica Bond platform. The slides were scanned and scored by 13 pathologists who estimated the percentage of malignant and immune cells expressing PD-L1. Statistical analyses were performed from December 1, 2015, to August 30, 2016, to compare antibodies and pathologists' scoring of tumor and immune cells.

MAIN OUTCOMES AND MEASURES Percentages of malignant and immune cells expressing PD-L1.

RESULTS Among the 90 samples, the SP142 assay was an outlier, with a significantly lower mean score of PD-L1 expression in both tumor and immune cells (tumor cells: 22c3, 2.96; 28-8, 3.26; SP142, 1.99; E1L3N, 3.20; overall mean, 2.85; and immune cells: 22c3, 2.15; 28-8, 2.28; SP142, 1.62; E1L3N, 2.28; overall mean, 2.08). Pairwise comparisons showed that the scores from the 28-8 and E1L3N tests were not significantly different but that the 22c3 test showed a slight (mean difference, 0.24-0.30) but statistically significant reduction in labeling of PD-L1 expression in tumor cells. Evaluation of intraclass correlation coefficients (ICCs) between antibodies to quantify interassay variability for PD-L1 expression in tumor cells showed high concordance between antibodies for tumor cell scoring (0.813; 95% CI, 0.815-0.839) and lower levels of concordance for immune cell scoring (0.277; 95% CI, 0.222-0.334). When examining variability between pathologists for any single assay, the concordance between pathologists' scoring for PD-L1 expression in tumor cells ranged from ICCs of 0.832 (95% CI, 0.820-0.844) to 0.882 (95% CI, 0.873-0.891) for each assay, while the ICCs from immune cells for each assay ranged from 0.172 (95% CI, 0.156-0.189) to 0.229 (95% CI, 0.211-0.248).

CONCLUSIONS AND RELEVANCE The assay using the SP142 antibody is an outlier that detected significantly less PD-L1 expression in tumor cells and immune cells. The assay for antibody 22c3 showed slight yet statistically significantly lower staining than either 28-8 or E1L3N, but this significance was detected only when using the mean of 13 pathologists' scores. The pathologists showed excellent concordance when scoring tumor cells stained with any antibody but poor concordance for scoring immune cells stained with any antibody. Thus, for tumor cell assessment of PD-L1, 3 of the 4 tests are concordant and reproducible as read by pathologists.

JAMA Oncol. 2017;3(8):1051-1058. doi:10.1001/jamaoncol.2017.0013
Published online March 9, 2017.

Invited Commentary page 1058

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: David L. Rimm, MD, PhD, Department of Pathology, Brady Memorial Laboratory 116, Yale University School of Medicine, 310 Cedar St, PO Box 208023, New Haven, CT 06520 (david.rimm@yale.edu).

Patient response to checkpoint inhibitor immunotherapy has been considered to be exceptional.[1-3] The checkpoint inhibitor programmed cell death ligand 1 (PD-L1) is the target for 1 therapy approved by the US Food and Drug Administration (FDA) (atezolizumab), and its receptor, programmed cell death 1 (PD-1), is the target for 2 others (nivolumab and pembrolizumab). In registrational trials, each of these drugs has been tested with a companion diagnostic assay that has been independently designed and is based on a combination of a unique antibody with a custom-designed assay using proprietary reagents, protocols, and thresholds defining elevated expression of PD-L1. This scenario has led to a challenge for pathologists who seek to provide companion diagnostic testing but do not necessarily know which therapeutic agent will be selected by the oncologist for any given patient.

Historically, immunohistochemistry (IHC) has been used to determine the presence or absence of a given protein. In combination with morphologic findings, IHC assists pathologists in classifying a tumor. Immunohistochemical assays are optimized by vendors to provide a binary outcome from what is inherently a continuous variable. Companion diagnostic tests are the exception to this approach for IHC since a threshold value is required. Expression above the threshold number of cells at or above the threshold intensity is then tightly linked to the prescription of a drug. The best examples of this testing are in breast cancer, in which estrogen receptor must be expressed in more than 1% of cells, at any intensity, to be considered positive.[4]

For PD-L1, 3 drug-specific tests are FDA approved as either companion (pembrolizumab) or complementary (atezolizumab and nivolumab) diagnostics, which use 3 different antibodies and 3 sets of assay conditions. The test for nivolumab uses the Dako/Agilent 28-8 assay, the test for pembrolizumab uses the Dako/Agilent 22c3 assay, and the test for atezolizumab uses the Ventana SP142 assay. This is a very different approach than that taken historically, in which, using the example of estrogen receptor, several common antibodies are used in either FDA-approved assays or laboratory-developed tests to give a result that can predict response to therapy for approximately 12 drugs that inhibit or otherwise modulate estrogen receptor–mediated signaling in breast cancer. This scenario raises a new problem for pathologists: specifically, whether they should be more concerned about accurate measurement of the target protein or focus on the assay result, as appears to now be required by the FDA in companion diagnostic testing for PD-L1, in which 3 separate assays are approved for the same protein.

This problem presents a theoretical issue and a practical issue. The first is: Does each of the assays equally assess the amount of PD-L1 present in the tissue? Although this is an important issue, the FDA does not require proof of the number of molecules expressed compared with some analytic standard. A more practical issue is: Are these FDA-approved assays equivalent as approved, and can any of the assays be used for any drug, or are the assays and prescribed scoring methods specific to the therapeutic with which they were developed? To address this practical question, 2 main efforts have begun to compare the assays in the United States. The first study, labeled the Blueprint study, is a comparison of 39 cases scored by 3 industry pathologists comparing the 3 approved tests as well as a fourth assay that is for investigational use only from AstraZeneca and Ventana based on the SP263 antibody.[5] This study showed concordance between 3 of the 4 assays, with the SP142 assay as an outlier. The Blueprint study was considered a pilot study and, as such, was not statistically powered nor was it multi-institutional.

The second United States–based study is reported here. Our study, sponsored by the National Comprehensive Cancer Network and funded by Bristol-Myers Squibb, sought to provide level 1 evidence for biomarker testing[6,7] by using a statistically powered, prospective design in a multi-institutional setting. The primary objective was to compare the performance of available antibodies, assays, and test platforms for the ability to accurately and reliably measure PD-L1. In the absence of patient outcome data across therapeutic products, we focused on direct comparison between 4 assays to understand the properties and performance of antibodies and tests relative to one another, evaluate differences in the assessment of PD-L1 on tumor cell surface vs immune infiltrates, and compare interpretation of results between pathologists across assays. Although limited by the use of samples from untreated patients, the level 1 evidence produced herein is not evidence for prediction or clinical accuracy of the assays but rather for assay concordance and pathologist concordance in the assessment of each assay.

Key Points

Question What is the concordance of 4 assays for programmed cell death ligand 1, including 2 assays cleared by the US Food and Drug Administration, 1 test for investigational use only, and 1 laboratory-developed test?

Findings This study found that 3 of the 4 assays were essentially equivalent, but 1 (SP142) identified only about 50% of patients who were positive for programmed cell death ligand 1 expression that were identified by the other 3 tests. Furthermore, scoring of the assays was highly concordant among pathologists for programmed cell death ligand 1 expression in tumor cells but not concordant for immune cells.

Meaning Three of the 4 assays (antibodies 22c3, 28-8, and E1L3N) appear to be interchangeable from an analytic perspective, but none of the assays has been clinically validated for cross-utilization.

Methods

Case Selection

A series of 90 surgically resected cases of non-small cell lung cancer (stages I-III), adenocarcinoma, and squamous cell carcinoma were obtained from the Yale School of Medicine Department of Pathology archives from January 1, 2008, to December 31, 2010 (eTable 1 in the Supplement). All tissue for the study was collected under Yale Human Investigation Committee Protocol No. 0304025173, which allows collection of tissue with consent or waiver of consent when no personalized health information is required, as was the case for this study.

Immunohistochemistry

Four 5-µm sections were cut from each case at Yale University and sent to 3 institutions for staining as follows: assay 1, to the University of Colorado for 22c3 on the Dako Link 48 platform; assay 2, to the University of Colorado for 28-8 on the Dako Link 48 platform; assay 3 to the Mayo Clinic, Rochester, New York, for SP142 on the Ventana Benchmark platform; and assay 4 to Yale University for Cell Signaling Technology E1L3N on the Leica Bond platform (as a laboratory-developed test). Although the Ventana assay is now FDA approved, it was slightly altered from the investigational use–only test that was available at the time of the staining for this study. That protocol is essentially identical to the current approved protocol with the exception of 3 steps representing different incubation times (eTable 2 in the Supplement). All other conditions were identical to the approved test, and the appearance of the slides is comparable to those using the approved protocol. For E1L3N, the staining procedure for the laboratory-developed test can be found in the eAppendix in the Supplement.

Pathologist Scoring

The stained slides were all sent to the University of Colorado for scanning by a Leica Aperio scanner and placement into a database viewable by using an internet connection. A template for scoring was constructed on the REDCap database. Pathologists scored the images conveyed via the internet, which allowed visualization of the entire slide with full zoom capacity from the equivalent of a 1 × original magnification to 400 × original magnification. Instructions were provided to 16 pathologists at 8 institutions, and a deadline was set for completion of scoring of 90 cases with 4 slides per case representing each pair of stain and platform. Despite extension of the deadline, only 13 of the 16 pathologists from 7 of the 8 institutions participating in the study correctly completed the scoring exercise.

Because each system has its own scoring protocol, we designed a unified scoring method for both tumor proportion scores (TPSs) and immune cell proportion scores (ICPSs) that could be used to calculate a score that fits into the categorical scoring system for each laboratory-developed test or FDA-approved assay, including the AstraZeneca Ventana SP263 test, even though that assay was not tested in this exercise. As in those assays, the score of the TPS or ICPS is based on membrane and cytoplasmic staining of any intensity. The scoring system is summarized in eTable 3 in the Supplement.

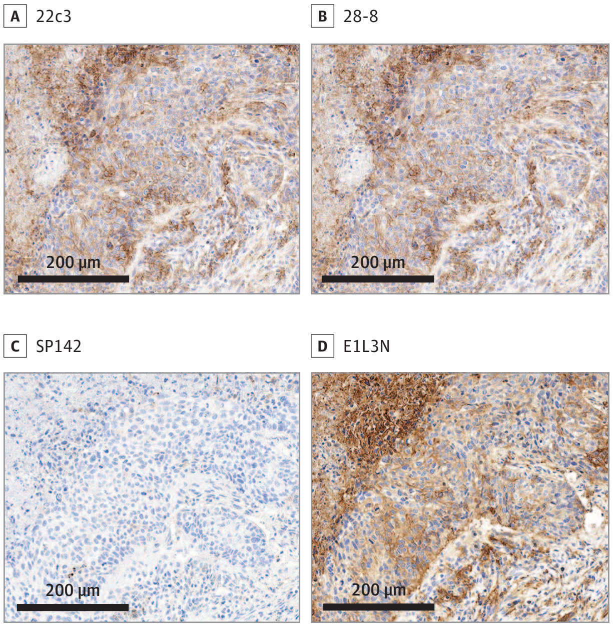
Statistical Analysis

Statistical analysis was performed from December 1, 2015, to August 30, 2016. Pathologists' scores were recorded on a 6-point scale, with each value corresponding to a range of the tumor percentage: the original score of category A is negative or less than 1% of tumor, a score of B is 1% to 4% of tumor, a score of C is 5% to 9% of tumor, a score of D is 10% to 24% of tumor, a score of E is 25% to 49% of tumor, and a score of F is 50% or more of tumor. The same statistical analyses were performed for the TPS and ICPS. For assay comparison, mean scores from the 13 pathologists were plotted for each antibody by cases. Paired Wilcoxon signed rank test was used to compare the antibody pairs for scores from individual pathologists. A mixed-effects linear model was used to evaluate the statistical significance in differences between antibodies, treating effects from pathologists as random effects. To assess concordance of the antibodies, intraclass correlation coefficients (ICCs) were calculated among the 4 antibodies and among 3 of the 4 tests (except SP142 on the Ventana Benchmark platform) using the mean scores of the pathologists for the 90 cases as well as each pathologist's scores. Sample size justification based on ICC was conducted before data collection. Assuming 4 antibodies, we calculated the statistical power that 90 slides can achieve to differentiate an almost perfect agreement (ICC, ≥0.85) from a moderate (ICC, 0.5) or strong (ICC, 0.7) agreement. Taking into account that approximately 35% of the scores will be positive, 90 slides can achieve 87.9% power at a significance level of $P < .05$ to differentiate an ICC of 0.85 from an ICC of 0.7. An ICC is interpreted as follows: below 0.3 indicates poor agreement, 0.5 indicates moderate agreement, 0.7 indicates strong agreement, and 0.85 or more indicates almost perfect agreement. Analyses were also performed to quantify the concordance of scores between the pathologists: ICC values between pathologists were calculated for each antibody in both the original 6 score levels and 3 aggregated levels (<1%, 1%-49%, and ≥50%).[8] Variance of the pathologists' scores was decomposed to contributions from antibodies and pathologists using analysis of variance. Furthermore, the original scores were dichotomized using the cutoff of greater than 50% and the cutoff of greater than 1% to assess the concordance between pathologists for binary tumor evaluation. The Fleiss κ coefficient and Kendall concordance coefficient were calculated to evaluate the agreement and concordance of the 13 pathologists' binary assessment for each antibody. A κ coefficient of 0.4 or less is poor to fair agreement, greater than 0.4 to 0.6 is moderate, greater than 0.6 to 0.8 is substantial, and greater than 0.8 is almost perfect. Strength of the Kendall concordance coefficient was interpreted similarly to that of the ICC.[9,10] Statistical analysis was completed using SAS software, version 9.4 (SAS Institute Inc) and MATLAB, version 2014b (The Mathworks Inc) based on the prescribed experimental design of the first phase of the National Comprehensive Cancer Network and Bristol-Myers Squibb study.

Results

Among the 90 samples, the appearance of the stained samples was similar to that seen previously in IHC assays of PD-L1 expression,[11-13] showing predominantly membranous staining. The 4 assays appeared largely similar, although 1 of the 4 assays was substantially lighter in staining intensity (**Figure 1**). **Figure 2**A and D show a comparison of the TPS and ICPS for each case by using the mean scores of the 13 pathologists as a continuous percentage score, even though each pathologist entered a categorical score, as shown in eTable 2 in the Supplement. Figure 2B, C, E, and F show the scoring results by

Figure 1. Immunohistochemical Images With Tumor Cell and Immune Cell Staining



The original magnification for all images is ×20.

percentages of patients in each categorical scoring class for each assay for both the TPS and ICPS, as well as the percentage positive using only the 50% and 1% cut points to generate a binary score for the TPS and using the 10% and 1% cut points for the ICPS.

To assess interassay variability, we first determined the mean score for the 13 pathologists for each antibody assay, then compared each antibody in pairwise comparisons to show the mean difference for each antibody, and then tested for significance using the Wilcoxon signed rank test and a mixed-effects model. **Table 1** shows the mean difference and statistical significance of each for both the TPS and ICPS. Only the 28-8 assay and the E1L3N assay were not statistically significantly different by this method, and the SP142 test had the greatest magnitude of difference compared with the other 3 antibody assays. The tumor means by assay are 22c3, 2.96; 28-8, 3.26; SP142, 1.99; and E1L3N, 3.20. The immune cell means by assay are 22c3, 2.15; 28-8, 2.28; SP142, 1.62; and E1L3N, 2.28. The difference in the means for the tumor cells that were significant were as follows: 22c3 was significantly lower than both 28-8 (mean difference, −0.3; $P < .001$) and E1L3N (mean difference, −0.246; $P < .001$). SP142 was significantly lower than all other assays as shown in Table 1. The ICC is perhaps a better method to compare these assays. Again using the mean of the 13 pathologists' scores, we found that the ICCs for the TPS and ICPS were 0.813 (95% CI, 0.815-0.839) and 0.277 (95% CI, 0.222-0.334), respectively, which increased to 0.971 and 0.804 when SP142 was excluded.

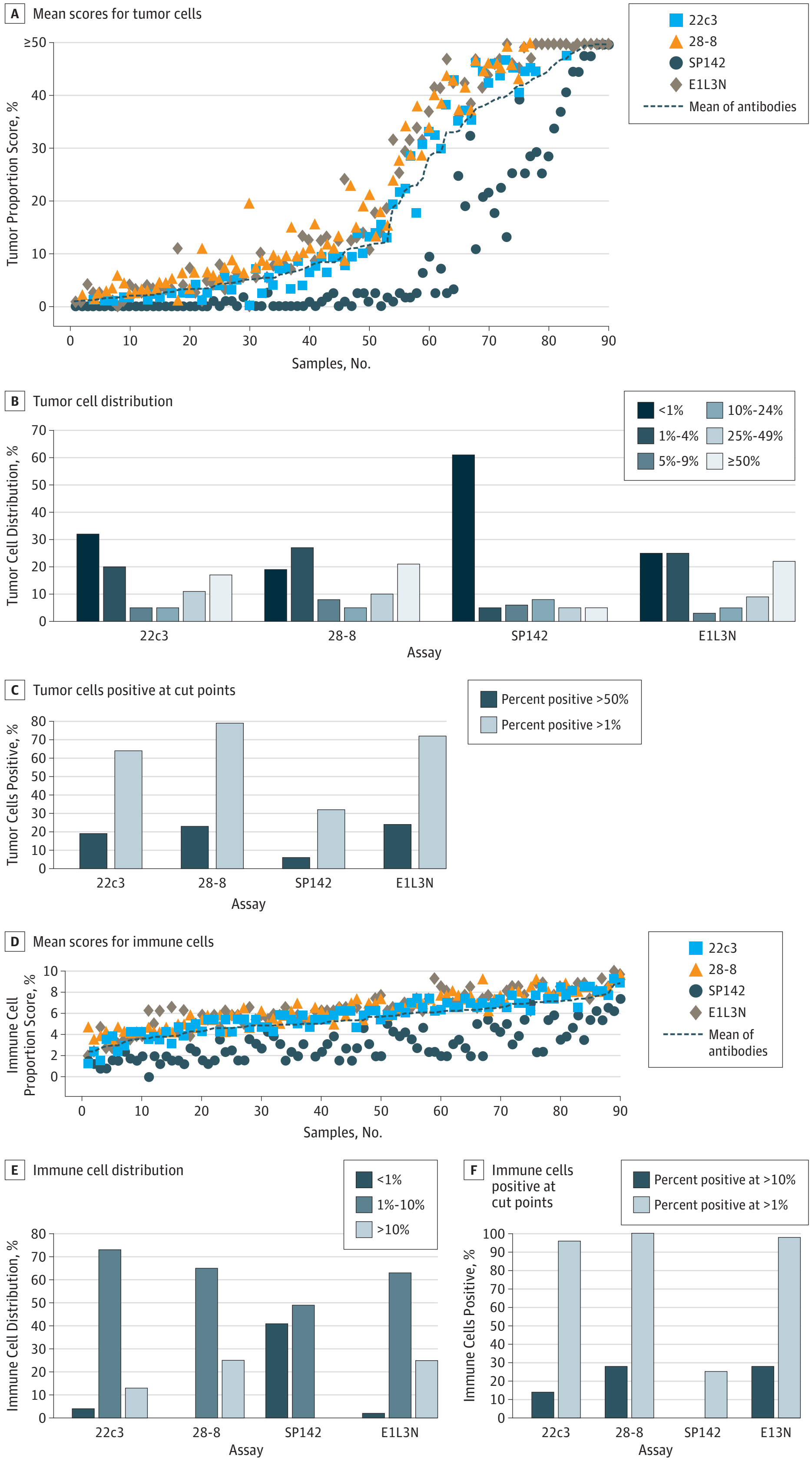
Although it is interesting to use the mean of the 13 pathologists' scores to compare the assays, the scoring of individual pathologists is more important since, in practice, a case is usually only examined by a single pathologist. The ICC for each pathologist and each antibody assay was measured to assess variability between pathologists in scoring both tumor and immune cells. **Table 2** shows the ICCs for each antibody assay for both tumor cell scoring and immune cell scoring. The concordance between pathologists' scores for tumor cells had an ICC of 0.882 (95% CI, 0.873-0.891) for 22c3, 0.832 (95% CI, 0.820-0.844) for 28-8, 0.869 (95% CI, 0.859-0.879) for SP142, and 0.859 (95% CI, 0.849-0.869) for E1L3N. In contrast, the ICCs for immune cells were markedly decreased, with an ICC of 0.207 (95% CI, 0.190-0.226) for 22c3, 0.172 (95% CI, 0.156-0.189) for 28-8, 0.185 (95% CI, 0.169-0.203) for SP142, and 0.229 (95% CI, 0.211-0.248) for E1L3N. A second important variable to determine for comparison of pathologists' scores is concordance around the cut point at which clinicians decide to prescribe drugs. At the time of this submission, there are FDA-approved cut points at greater than 50% and greater than 1%. The concordance, as measured by the Fleiss κ statistic for the mean of all 4 antibody assays at the cut point of greater than 50%, is 0.749 and at the cut point of greater than 1% is 0.537. The Kendall concordance for the mean of all 4 antibody assays at the cut point of greater than 50% is 0.775 and at the cut point of greater than 1% is 0.612. Our study does not have outcome information for anti-PD-1 or anti-PD-L1 therapies. As such, the sensitivity and specificity of the assay could not be determined. However, in efforts to evaluate the ability of any given pathologist to correctly assess each assay, we defined the median pathologist's score as "truth" and calculated the correctly predicted proportion of positive cases as an analogue for sensitivity and a correctly predicted proportion negative as an analogue for specificity. **Figure 3** shows these statistics as each of 3 possible cut points: greater than 1%, greater than 5%, and greater than 50%.

Discussion

The SP142 assay was associated with statistically significantly lower levels of PD-L1 staining than the other 3 assays for both the TPS and ICPS. The 22c3 assay also showed statistically significantly lower levels of PD-L1 expression compared with both the 28-8 and E1L3N assays, but this slightly lower level of PD-L1 staining was detected when only a mean of the 13 pathologists' scores was used. Also, we found that pathologists were highly concordant for each assay, with ICCs of approximately 0.8 for the TPS across any single assay, but poorly concordant for the ICPS, with ICCs of approximately 0.2. This finding suggests that IHC may be a good method for assessment of PD-L1 in tumor cells but is probably inadequate for assessment of immune cell expression independent of which assay is selected. In tumor cells, we found higher concordance at the 50% cut point than at the 1% cut point. The 1% cut point may require the use of automated systems or training regimens for pathologists to improve assay precision.

Because we used a unified scoring system, it allowed us to assess the pathologists' ability to score at various TPS levels. The absence of "truth" or data on response to therapy limits our observations, but definition of a surrogate for "truth,"

Figure 2. Tumor Proportion Scores and Immune Cell Proportion Scores



A, Mean scores for tumor cells from all 13 pathologists for each of the 90 slides. B, Frequency distributions for tumor cells. C, Percentage of tumor cells positive at the greater-than-50% and greater-than-1% cut points. D, Mean scores for immune cells from all 13 pathologists for each of the 90 slides. E, Frequency distributions for immune cells. F, Percentage of immune cells positive at the greater-than-10% and greater-than-1% cut points.

Table 1. Pairwise Assay Comparison

	Tumor Cell Score			Immune Cell Score		
Antibody Pair	Mean (SD)[a]	*P* Value[b]	Mixed Effects *P* Value[c]	Mean (SD)[a]	*P* Value	Mixed Effects *P* Value
22c3 and 28-8	-0.300 (0.393)	<.001	<.001	-0.127 (0.164)	<.001	.001
22c3 and SP142	0.970 (1.000)	<.001	<.001	0.535 (0.288)	<.001	<.001
22c3 and E1L3N	-0.246 (0.372)	<.001	.003	-0.128 (0.189)	<.001	.002
28-8 and SP142	1.270 (1.081)	<.001	<.001	0.662 (0.294)	<.001	<.001
28-8 and E1L3N	0.055 (0.415)	.21	.28	-0.001 (0.194)	.96	.97
SP142 and E1L3N	-1.216 (1.121)	<.001	<.001	-0.664 (0.333)	<.001	<.001

[a] Mean of 13 pathologists.

[b] Wilcoxon signed rank test.

[c] Paired *t* test incorporating random effects of pathologists.

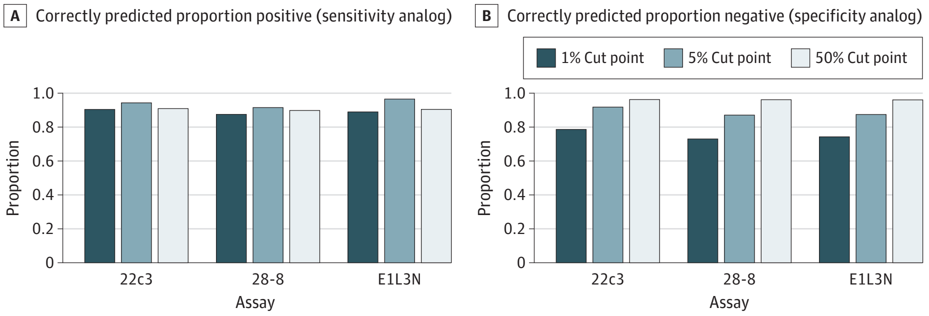
Table 2. ICC for the Pathologist Scores and Concordance Statistics

	Antibody, ICC (95% CI)				
Cells[a]	22c3	28-8	SP142	E1L3N	Summary, Mean (SD)
Tumor cells	0.882 (0.873-0.891)	0.832 (0.820-0.844)	0.869 (0.859-0.879)	0.859 (0.849-0.869)	0.86 (0.02)
Immune cells	0.207 (0.190-0.226)	0.172 (0.156-0.189)	0.185 (0.169-0.203)	0.229 (0.211-0.248)	0.19 (0.03)

Abbreviation: ICC, intraclass correlation coefficient.

[a] N = 90.

Figure 3. Proportion of Correctly Predicted Positive and Negative Cases



A, Correctly predicted proportion of positive cases by cut point.
B, Correctly predicted proportion of negative cases by cut point.

the median pathologists' score, allowed us to further dissect where pathologists agreed and where they did not in a more real-world manner. In companion diagnostic tests, high assay sensitivity is required for the identification of every patient that may benefit. This approach favors a lower cut point to increase the percentage of patients who are treated. However, if too many patients who are predicted by the test to respond do not respond, either the test, the drug, or both are more likely to fail. As such, we have generated a surrogate for sensitivity by calculating the percentage of times a single pathologist would call the test positive if he or she exceeded the median score of all pathologists at each cut point. These data show that, as designed, these assays as scored by our pathologist group have a 90% to 95% sensitivity for any of the tested cut points to predict a positive test result. However, we also used the same approach to see the proportion of pathologists who are lower than the median score. This surrogate for specificity shows that, for each assay, the 1% cut point has between 70% and 80% specificity compared with greater than 90% for the 5% cut point and greater than 95% for the 50% cut point. Although this is only a theoretical estimate of the potential sensitivity and specificity, the model shows that high specificity requires a high cut point and that high sensitivity can be obtained across all thresholds.

Limitations

A key limitation of this effort is the lack of outcome data since these patients were not treated with PD-1 or PD-L1 axis therapies. As such, we can only evaluate this work in the context of assay comparisons and not clinical concordance. However, the distribution of PD-L1 expression at the 50% and 1% cut points closely reflected the percentages of the population considered positive in the Keynote[14] and CheckMate[15] studies. This study is a comparison of the assays as performed in 3 specific laboratories using the best possible practices. Recently, it was shown that the antibodies, from the perspective of interaction with the PD-L1 epitope, are most likely only subtly different, if at all.[16] Thus, the variation seen in our study is most likely a function of the recipe or protocol for each assay. Therefore, another limitation of this study is that the assay used for SP142 on the Ventana platform is not identical to the platform now approved by the FDA. We believe the difference is minimal and note the similarity in the appearance of the images seen in our study to those shown in the Blueprint study.[5,17] Furthermore, the differences in the assays (eTable 2 in the Supplement) appear to be minimal. However, because the solutions are proprietary, we cannot exclude the possibility that these small differences result in large effects and are the cause of the lower levels of expression seen in our study.

Conclusions

This study represents level 1 evidence for the comparison of these biomarkers. We have shown that the SP142 assay is an outlier and that pathologists are much better at scoring the TPS than the ICPS. There appears to be minimal difference between the other 3 assays tested, which could have implications for assay choices in individual pathologists' laboratories where there is financial pressure to validate only a single PD-L1 assay. We hope that these observations will lead to future clinical concordance studies in patients treated with PD-1 axis therapies.

ARTICLE INFORMATION

Accepted for Publication: December 23, 2016.

Published Online: March 9, 2017. doi:10.1001/jamaoncol.2017.0013

Open Access: This article is published under *JAMA Oncology's* open access model and is free to read on the day of publication.

Author Affiliations: Department of Pathology, Yale University School of Medicine, New Haven, Connecticut (Rimm, Homer); Department of Epidemiology and Biostatistics, Texas A&M University School of Public Health, College Station (Han); Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland (Taube, Anders); Department of Anatomic Pathology, Mayo Clinic, Rochester, Minnesota (Yi, Roden); Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha (Bridge, West); Department of Pathology, Fox Chase Cancer Center, Philadelphia, Pennsylvania (Flieder, Wu); Department of Translational Molecular Pathology, The University of Texas M.D. Anderson Cancer Center, Houston (Fujimoto, Wistuba); Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora (Yu, Kowalewski, Rivard, Hirsch); Department of Pathology, NorthShore University Health System, Evanston, Illinois (Rehman); Department of Immuno-Oncology, Bristol-Myers Squibb, Plainsboro, New Jersey (Batenchuk, Burns).

Author Contributions: Drs Rimm and Wistuba had full access to all the data in the study, and Dr Rimm takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Rimm, Han, Taube, Yi, Bridge, Hirsch, Wistuba.
Acquisition, analysis, or interpretation of data: All authors.
Drafting of the manuscript: Rimm, Han, Taube, West, Fujimoto, Kowalewski, Hirsch.
Critical revision of the manuscript for important intellectual content: Rimm, Han, Yi, Bridge, Flieder, Homer, West, Wu, Roden, Yu, Anders, Rivard, Rehman, Batenchuk, Burns, Hirsch, Wistuba.
Statistical analysis: Rimm, Han, Rehman.
Obtained funding: Burns, Wistuba.
Administrative, technical, or material support: Han, Yi, Bridge, Homer, West, Wu, Fujimoto, Yu, Anders, Kowalewski, Rivard, Burns, Hirsch.
Study supervision: Rimm, Yi, Flieder, Rivard, Batenchuk, Hirsch, Wistuba.

Conflict of Interest Disclosures: Dr Rimm reported serving as a consultant for AstraZeneca, Agendia, Bethyl Labs, Biocept, Bristol-Myers Squibb, Cell Signaling Technology, ClearSight, Genoptix/Novartis, Merck, OptraScan, Perkin Elmer, and Ultivue and receiving research grant support from Cepheid, Genoptix, Gilead Sciences, Pierre Fabre, and Perkin Elmer. Dr Taube reported serving as a consultant for Bristol-Myers Squibb, AstraZeneca, and Merck and receiving research funding from Bristol-Myers Squibb. Dr Bridge reported serving as a consultant for Cepheid Inc and National Comprehensive Cancer Network–Peregrine. Dr Anders reported serving on the advisory board of Adaptive Biotechnologies and receiving research support from Bristol-Myers Squibb, Five Prime Diagnostics, National Institutes of Health, Stand Up 2 Cancer, and Fibrolamellar Cancer Foundation. Dr Hirsch reported serving on the advisory boards of Genentech/Roche, Bristol-Myers Squibb, Merck, AstraZeneca, Novartis, Pfizer, Ventana, and HTG Molecular Diagnostics Inc. Dr Wistuba reported receiving honoraria from Genentech/Roche, Bristol-Myers Squibb, Boehringer Ingelheim, AstraZeneca, Pfizer, Ariad, HTG Molecular Diagnostics Inc, and Asuragen and receiving research grants from Genentech/Roche, Bristol-Myers Squibb, HTG Molecular Diagnostics Inc, and Asuragen. Drs Batenchuk and Burns are employed by Bristol-Myers Squibb.

Funding/Support: This study was funded by Bristol-Myers Squibb in collaboration with the National Comprehensive Cancer Network Oncology Research Program.

Role of the Funder/Sponsor: The funding sources played a role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication and as such are included in the author list and acknowledgments as per Bristol-Myers Squibb and National Comprehensive Cancer Network policy.

Additional Contributions: Kimary Kulig, PhD, MPH, Bristol-Myers Squibb, participated in the concept and design of the study. Patricia Esposito, MPH, National Comprehensive Cancer Network, Oncology Research Program, and Nabil Chehab, PhD, MBA, Bristol-Myers Squibb, assisted in the execution of the study. Kulig and Chehab were not compensated for their contributions. Esposito was compensated for her contribution.

REFERENCES

1. Brahmer JR, Tykodi SS, Chow LQ, et al. Safety and activity of anti–PD-L1 antibody in patients with advanced cancer. *N Engl J Med*. 2012;366(26):2455-2465.

2. Topalian SL, Hodi FS, Brahmer JR, et al. Safety, activity, and immune correlates of anti–PD-1 antibody in cancer. *N Engl J Med*. 2012;366(26):2443-2454.

3. Herbst RS, Soria JC, Kowanetz M, et al. Predictive correlates of response to the anti–PD-L1 antibody MPDL3280A in cancer patients. *Nature*. 2014;515(7528):563-567.

4. Hammond ME, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Arch Pathol Lab Med*. 2010;134(6):907-922.

5. Hirsch FR, McElhinny A, Stanforth D, 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(2):208-222.

6. Hayes DF, Bast RC, Desch CE, et al. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J Natl Cancer Inst*. 1996;88(20):1456-1466.

7. Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*. 2009;101(21):1446-1452.

8. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284-290. doi:10.1037/1040-3590.6.4.284

9. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174.

10. Legendre P. Species associations: the Kendall coefficient of concordance revisited. *J Agric Biol Environ Stat*. 2005;10(2):226-245. doi:10.1198/108571105X46642

11. Phillips T, Simmons P, Inzunza HD, et al. Development of an automated PD-L1 immunohistochemistry (IHC) assay for non–small cell lung cancer. *Appl Immunohistochem Mol Morphol*. 2015;23(8):541-549.

12. Taube JM, Anders RA, Young GD, et al. Colocalization of inflammatory response with B7-h1

expression in human melanocytic lesions supports an adaptive resistance mechanism of immune escape. *Sci Transl Med*. 2012;4(127):127ra37.

13. Velcheti V, Schalper KA, Carvajal DE, et al. Programmed death ligand-1 expression in non–small cell lung cancer. *Lab Invest*. 2014;94(1):107-116.

14. Garon EB, Rizvi NA, Hui R, et al; KEYNOTE-001 Investigators. Pembrolizumab for the treatment of non–small-cell lung cancer. *N Engl J Med*. 2015;372(21):2018-2028.

15. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer. *N Engl J Med*. 2015;373(17):1627-1639.

16. Gaule P, Smithy JW, Toki M, et al. A quantitative comparison of antibodies to programmed cell death 1 ligand 1 [published online August 18, 2016]. *JAMA Oncol*. doi:10.1001/jamaoncol.2016.3015

17. Ratcliffe MJ, Sharpe A, Midha A, et al. Agreement between programmed cell death ligand-1 diagnostic assays across multiple protein expression cut-offs in non-small cell lung cancer [published online January 10, 2017]. *Clin Cancer Res*. 2017;clincanres.2375.2016. doi:10.1158/1078-0432.CCR-16-2375

Invited Commentary

Assays for PD-L1 Expression Do All Roads Lead to Rome?

Gabriel L. Sica, MD, PhD; Suresh S. Ramalingam, MD

Related article page 1051

Inhibition of the programmed cell death 1 protein (PD-1) pathway reverses T-cell exhaustion and improves survival relative to standard chemotherapy for patients with advanced stages of non-small cell carcinoma (NSCLC).[1-3] Since the clinical benefit is restricted to a subset of patients, predictive biomarkers are essential for patient selection. A number of putative markers have demonstrated predictive potential, but the only proven marker to date is expression of the PD-1 ligand (PD-L1), assessed by immunohistochemistry (IHC).[4]

In advanced NSCLC, the biomarker panel includes testing for *EGFR* (OMIM 131550) mutation and rearrangements in the *ALK* (OMIM 105590) and *ROS1* (OMIM 165020) genes. Recently, PD-L1 expression has been added to the biomarker panel to select patients for immune checkpoint inhibitor therapy. Consequently, the role of the pathologist in providing accurate diagnosis and biomarker testing has become even more important in the diagnostic workup of lung cancer. Tissue availability is a major limiting factor in optimal biomarker testing for lung cancer since tumor biopsies from patients with lung cancer are often limited in quantity owing to the biopsy type and accessibility of the tumor to biopsy. At least 4 anti–PD-1 or anti–PD-L1 inhibitors, each with its own companion or complementary diagnostic test kit, are approved for either a Dako-based (nivolumab and 28-8; pembrolizumab and 22c3) or Ventana-based (atezolizumab and SP142; durvalumab and SP263) IHC staining platform. With the exception of durvalumab, all the drugs are approved by the US Food and Drug Administration for the treatment of advanced NSCLC. In addition to the use of different antibodies, each of the PD-L1 tests has disparate guidelines for interpreting IHC staining. This situation further complicates the ancillary testing algorithm regarding which PD-L1 test to perform unless the pathologist knows the specific drug that will be used for treatment, and thereby would preclude reflex testing for PD-L1 on diagnosis of lung cancer. The lack of harmonized testing and interpretation for PD-L1 also has the potential for increased tissue use if multiple assays are requested and may cause a delay in the return of results. In addition, only the largest reference laboratories will have the resources and incentive to validate all 4 tests.

Harmonization of the PD-L1 assays would have multiple benefits, including the establishment of the interchangeability of the various assays and requirement for validation of just one type of test. A major step in the harmonization of the PD-L1 assays is to determine whether the reagents are equivalent with regard to detection of PD-L1, and whether the individual assays will be interchangeable to identify patients who may derive clinical benefit.

In an article in this issue of *JAMA Oncology* with direct clinical relevance, Rimm and colleagues[5] compared PD-L1 IHC staining profiles in 90 surgically resected NSCLC specimens using 4 different monoclonal antibodies (Dako 28-8, Dako 22c3, Ventana SP142, and Cell Signaling Technology E1L3N) on 3 different IHC platforms (Dako Link 48 Platform, Ventana Benchmark Platform, and Leica Bond Platform). Of these 4 IHC assays, only the E1L3N assay is a laboratory-derived test. In addition to determining the characteristics of the 4 assays, the study also investigated the reproducibility of 13 pathologists' scoring of the tumor and immune cell infiltrates. Of the 4 assays, the Ventana SP142 was a significant outlier, with lower levels of PD-L1 detection in tumor cells. These results are similar to those in another report by Hirsch et al[6] (the Blueprint study) that analyzed 38 cases of NSCLC using 4 PD-L1 IHC assays (Dako 28-8, Dako 22c3, Ventana SP142, and Ventana SP263). Rimm et al[5] also noted that the SP142 assay detected only half as many PD-L1 cases that were highly positive (>50%) in tumor cells compared with the 3 other assays tested. Similar results were obtained when evaluating the immune cell compartment for PD-L1 expression. Assessment of the interassay variability, based on mean scores from all 13 pathologists, showed that only 2 assays, 28-8 and E1L3N, showed equivalence, while both 22c3 and SP142 showed statistically significant variable results, with SP142 showing the highest level of variation and 22c3 showing only a slight variation. When intraassay variation between the pathologists was assessed, there was high concordance for percentage of tumor staining but not for the immune cell compartment. The