


REVIEW

Challenges facing pathologists evaluating PD-L1 in head & neck squamous cell carcinoma

Ilaria Girolami¹ | Liron Pantanowitz² | Massimo Barberis³ | Gaetano Paolino⁴ |
Matteo Brunelli⁴ | Elena Vigliar⁵ | Enrico Munari⁶ | Swati Satturwar⁷ |
Giancarlo Troncone⁵ | Albino Eccher⁴ 

¹Division of Pathology, Central Hospital Bolzano, Bolzano, Italy

²Department of Pathology & Clinical Labs, University of Michigan, Ann Arbor, MI, USA

³Division of Pathology, IEO European Institute of Oncology, Milan, Italy

⁴Department of Pathology and Diagnostics, University and Hospital Trust of Verona, Verona, Italy

⁵Department of Public Health, University of Naples Federico II, Naples, Italy

⁶Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy

⁷Department of Pathology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

Correspondence

Albino Eccher, Department of Pathology and Diagnostics, University and Hospital Trust of Verona, P.le Stefani n. 1; 37126, Verona, Italy.

Email: albino.eccher@aovr.veneto.it

Abstract

Background: Programmed death-ligand 1 (PD-L1) expression with combined positive score (CPS) ≥ 1 is required for administration of checkpoint inhibitor therapy in recurrent/metastatic head and neck squamous cell carcinoma (HNSCC). The 22C3 pharmDx Dako immunohistochemical assay is the one approved as companion diagnostic for pembrolizumab, but many laboratories work on other platforms and/or with other clones, and studies exploring the potential interchangeability of assays have appeared.

Evidence from the literature: After review of the literature, it emerges that the concordance among assays ranges from fair to moderate, with a tendency of assay SP263 to yield a higher quota of positivity and of assay SP142 to stain better immune cells. Moreover, pathologists achieve very good concordance in assessing PD-L1 CPS, particularly with SP263.

Conclusions: Differences in terms of platforms, procedures, and study design still preclude a quantitative synthesis of evidence and clearly further work is needed to draw stronger conclusions on the interchangeability of PD-L1 assays in HNSCC.

KEYWORDS

combined positive score, head and neck squamous cell carcinoma, immunohistochemical assay, programmed death-ligand 1, review

1 | INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC) affects approximately 880 000 new patients each year worldwide and represents a leading cause of mortality in some countries.¹ Despite combined therapy that includes surgical resection with radiotherapy and/or chemotherapy, the 5-year overall survival has improved only modestly over the past three decades and is only 50%–65%.^{2,3} A turning point in the therapy of HNSCC ensued with the introduction of immunotherapy targeting the programmed death-1 (PD-1)/programmed death-ligand-1

(PD-L1) axis.^{4–6} The US Food and Drug Administration (FDA) and European Medicines Agency (EMA) approved of the PD-1 inhibitors pembrolizumab and nivolumab for recurrent and metastatic disease. The administration of these drugs showed improved survival with reduced toxicity.^{7–10} Combination chemotherapy with other PD-1/PD-L1 inhibitors atezolizumab, avelumab, cemiplimab, and durvalumab is currently being tested in clinical trials,⁵ and interest is growing also for chemopreventive application.^{11,12} However, the therapeutic benefit of single patients being treated with immune checkpoint inhibitors (ICI) is often not as impressive as expected.

Programmed death-ligand 1 expression in tumor and immune cells has been evaluated in clinical trials in relation to therapeutic response.¹³ PD-L1 expression can be assessed by means of immunohistochemistry (IHC) with a tumor proportion score (TPS), which is the proportion of positive tumor cells, or with a combined positive score (CPS), which is defined by the ratio of total positive tumor and immune to the total number of viable tumor cells. These scoring systems are variably applied in clinical trials. In the KEYNOTE-040 study, patients with recurrent or metastatic HNSCC treated with pembrolizumab showed significantly improved survival when their tumor biopsies expressed PD-L1 with TPS $\geq 50\%$.⁷ PD-L1 expression showed the best predictive performance when using a CPS with a cutoff of ≥ 20 , and a lower but significant benefit in survival when CPS was ≥ 1 for first-line treatment in the KEYNOTE-048 study.⁵ Other studies also consider a cutoff of $\geq 1\%$ for both TPS and CPS as clinically relevant.⁹ Recently, post hoc analysis of the KEYNOTE-040 trial showed that CPS ≥ 50 is equivalent to TPS $\geq 50\%$ for predicting objective response rate, overall survival, and progression-free survival in HNSCC patients.¹⁴ Differences in prediction of response with the two scoring systems were predicted to be tied to inclusion of immune cells in the CPS, as around 50%–60% of HNSCC tumor cells express PD-L1 when assessed with TPS, but this percentage increases to 85% when considering both tumor and surrounding immune cells, as is measured with the CPS.^{13,15} Consequently, CPS appears to be more sensitive at lower cutoffs of positivity, supporting the importance of PD-L1-positive immune cells.¹⁴ Finally, international agencies included a CPS of ≥ 1 as a selection criterion for first-line treatment of recurrent and metastatic HNSCC with pembrolizumab. However, similarly to what transpired in lung cancer, there are different PD-L1 IHC assays, using different PD-L1 antibodies (22C3, 28-8, SP263, SP142, E1L3N), on different IHC platforms.

Varying antibody clones and platforms have been approved for each available PD-1 and PD-L1 inhibitor, making comparison among immunotherapy trials difficult.¹⁶ Two of these clones, 22C3 and 28-8, both run on the Dako immunohistochemistry platform, are the approved companion and complementary diagnostic test for pembrolizumab and nivolumab, respectively.^{17,18} However, many institutions routinely work on other platforms (e.g., Roche), creating a problem as they are accordingly not then able to offer these originally approved assays. A laboratory developed test (LDT) is any test differing from the original regulatory approved commercial PD-L1 clone assay, no matter how small the difference is to one or more of its components/procedures.¹⁹ Unfortunately, LDT development can be limited by an inability to standardize many of the assay components. As a result, LDTs are likely to be less robust than commercial tests and may thus introduce variability to results.¹⁹ As occurred previously with lung cancer, harmonization studies among different assays and studies were necessary to compare their diagnostic performance. Herein, we review the current field regarding the assessment of PD-L1 in HNSCC, compare these

varied assays, and underscore challenges facing pathologists required to interpret PD-L1 expression.

2 | METHODS

The review question was modeled on a Population, Intervention/Index, Comparator, Outcome (PICO) model. Population was represented by HNSCC only cases, Index test was any PD-L1 IHC clone on any platform, and Comparator term was considered any other clone or platform deployed to evaluate PD-L1 staining. We were interested in two outcomes: the concordance among different assays and secondarily the concordance among reading pathologists with any of the assays. A systematic search was carried out in Pubmed and Embase electronic database until January 10, 2021, with the combination of the key terms “PD-L1” and “HNSCC” defined with all their aliases. We did not insert an Outcome term in the search strategy to keep it as broad as possible. The complete search strategy is found in, Table S1. Inclusion criteria were the presence of any type of comparison among two or more assays for IHC PD-L1 expression in HNSCC or the comparison among pathologists assessing PD-L1 with one or more assays and the presence of reported concordance measures of any type (intraclass correlation coefficient (ICC), percent agreement, Cohen's kappa and variants) as outcomes of interest. No language or type of paper restrictions were applied. Studies not dealing with HNSCC (wrong Population), not dealing with PD-L1 assessment in IHC (wrong Index), or not presenting any type of comparison were excluded.

Quality of studies was assessed with a modified QUADAS2 tool tailored on our review question. We removed the question on case-control study design as it did not pertain to our review question, and we added the question on presence of detailed technical procedure with clone and platform. The modified set of items is found in Table S2.

3 | RESULTS

Of 3379 items after removal of duplicates, after screening according to title and abstract, 180 were assessed in full-text form and 19 were considered informative and included. The flow of article screening is found in Figure 1. The studies were represented by 10 full articles and 9 abstracts. The studies dealt with HNSCC cases from USA and Canada ($n = 10$, 53%), Europe ($n = 7$, 37%), Israel ($n = 1$, 5%), and Latin America ($n = 1$, 5%). The clone 22C3 was used in 14 (74%) studies, clone SP263 in 13 (68%), clone SP142 in 10 (53%), clone 28-8 in 5 (26%), and clone E1L3N in 2 (11%). Eight studies specifically evaluated PD-L1 expression with CPS, while the others evaluated separately tumor and immune cells with different cutoffs. The summary of the studies is found in Table 1, while full list of references is found in Supplementary material Appendix.

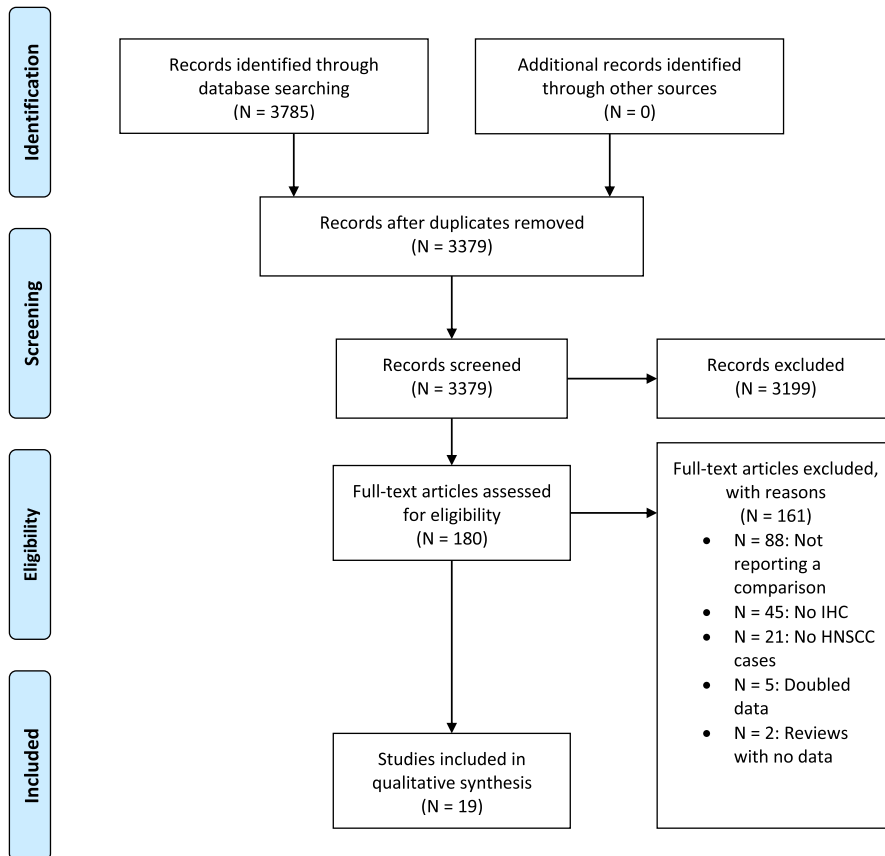


FIGURE 1 Flow of article screening according to PRISMA

The summary of quality appraisal is reported in Figure S1. The main item with risk of bias was the selection of cases, given that often it was not specified whether the cases were randomly or consecutively selected. Moreover, a great quota of studies was represented by abstracts with limited reporting, thus leading to a high number of studies with unclear risk of bias.

Given the high heterogeneity of studies, it was not possible to perform a meta-analysis; hence, we provide an evidence-based review with qualitative discussion of most relevant studies, dividing the discussion according to the main outcomes: the concordance among assays and among pathologists.

3.1 | Concordance among assays

Compared to lung cancer, evidence for direct comparison among assays in HNSCC is scarce. De Ruiter et al.²⁰ compared the reference 22C3 assay with a 22C3 laboratory developed test (LDT) and the SP263 assay in 143 cases. A smaller series published by Crosta et al.²¹ tested the performance of five different PD-L1 protocols with clones SP142, SP263, and 22C3 on various platforms against the 22C3 pharmDx on 15 cases/30 cores. The aforementioned studies utilized a tissue microarray (TMA) and were comprised of cases from several sites in the head and neck region, with PD-L1 assessed centrally by trained pathologists. These studies both applied clinically relevant cutoffs of CPS =1 and =20 which emerged from clinical trials, while previous studies were based on TPS with various

cutoffs, with or without distinction between staining of tumor or immune cells. Considering the quota of cases that are labeled as positive with CPS1-20 or higher than 20, both studies point toward an increased positivity rate when using the SP263 Ventana assay in comparison with the reference 22C3 pharmDx assay. The SP263 assay showed the greatest quota of positive cases compared with the other assays as reported by Crosta et al., while the 22C3 assay on the Omnis Dako platform showed the lowest. Crosta et al. also evaluated the performance of the various assays against the reference test by means of sensitivity and specificity for positive CPS ≥ 1 and found that the SP142 clone on the Ventana Benchmark Ultra platform performed best with a sensitivity of 92% and 100% specificity. The sensitivity and specificity of the other 22C3 assays run on other platforms ranged from 79% to 88% sensitivity and 80% to 83% specificity, while the SP263 assay showed 96% sensitivity and 50% specificity. This implies there is either a high quota of false positive results or that cases labeled as negative with the reference assay were incorrectly evaluated as positive with the SP263 assay. However, the negative predictive value of the best-performing SP142 was only 67%, meaning that one out of three negative cases with SP142 is not truly negative with 22C3 pharmDx. However, the Crosta et al. study has some limitations, such as the very limited number of cases evaluated, notable quota of non-evaluable cases was included, and concordance measures stronger than simple percent agreement such as ICC and kappa were not reported for pairwise comparisons. On the other hand, De Ruiter et al. reported precisely on the ICC and Cohen's kappa for their three comparisons and found

TABLE 1 Summary of retrieved studies dealing with comparison among assays in HNSCC

Author, year (Country)	N cases	Sites	Scoring system	Clone	Main findings	Main limitations/Notes
Al-Masri, 2017 (USA)	200–500	HNSCC NOS	TC and IC, NOS	<ul style="list-style-type: none"> SP142 Ventana SP263, Ventana 22C3, DAKO 28–8, DAKO 	<ul style="list-style-type: none"> Strong intrinsic agreement in TC and IC PD-L1 expression between the SP263 and 22C3 and 28–8 assays. IC scoring more variable than TCs SP142 assay with lower agreement versus the other assays in both NSCLC and HNSCC 	<ul style="list-style-type: none"> Abstract only, no quantitative measures reported Also NSCLC cases
Crosta, 2021 (Italy)	15 (30 cores)	OC, P, L	CPS with 1 and 20 cutoffs	<ul style="list-style-type: none"> 22C3 pharmDx Dako Autostainer as gold standard 22C3 Ventana Benchmark SP142 Ventana Benchmark 22C3 Leica Bond SP263 Ventana Benchmark 22C3 Dako Omnis 	<ul style="list-style-type: none"> Best performance with the SP142 clone on the Ventana Benchmark Ultra platform (92% and 100% sensitivity and specificity) Sensitivity (88%–96%) and specificity (50%–100%) varying among protocols of staining Higher proportion of positive cases with SP263 (sensitivity 96%, specificity 50%) ICC 0.774 for interobserver reliability and fair to moderate agreement for the cutoffs 	<ul style="list-style-type: none"> Few cases and some selection bias Great quota of unevaluable cases Concordance measures are not reported for pairwise comparisons
De Meulenaere, 2017 (Belgium)	99	OP	Positive if $\geq 5\%$ TC or IC	<ul style="list-style-type: none"> 22C3 Agilent-Dako 1:100 with Ultraview kit Ventana SP142 Roche with Optiview kit Ventana 	<ul style="list-style-type: none"> More TC positive with 22C3 than SP142 (34% vs. 23%) Low levels of positivity in IC (2%–3%) Positivity in TC strongly linked with positivity in immune environment with both clones OS associated with PD-L1 expression with SP142 clone only 	<ul style="list-style-type: none"> No use of scoring systems used in trials (TPS or CPS) No direct comparison among clones
De Meulenaere, 2018 (Belgium)	99	OP	Cutoffs 1%, 5%, 10%	<ul style="list-style-type: none"> 22C3 Agilent-Dako 1:100 with Ultraview kit Ventana SP142 Roche with Optiview kit Ventana 	<ul style="list-style-type: none"> Moderate agreement (κ 0.511) among clones and OPA 75%–83.3% at various cutoffs Substantial agreement among four scoring pathologists (ICC >0.70) for both clones OPA among pathologists at various cutoffs 66%–81% for SP142 and 65%–74% for 22C3 Strong correlation between manual and digital assessment of PD-L1 for both clones 	<ul style="list-style-type: none"> No special limitations; the study explores also the agreement between different type of specimens (resection vs biopsy, primary vs node metastasis), finding that biopsy underscores PD-L1 expression with both clones
De Ruiter, 2020 (Netherlands)	147	OP, HP, L	TPS and CPS	<ul style="list-style-type: none"> 22C3 pharmDx assay on the Dako Link 48 SP263 assay on the Ventana Benchmark Ultra 22C3 as an LDT on Ventana Benchmark Ultra 	<ul style="list-style-type: none"> ICC always lower than moderate (<0.50) and κ in the range of poor or fair for the relevant cutoffs among SP263 and 22C3 pharmDx for both TPS and CPS For CPS, concordance of 22C3 LDT with the SP263 assay and with the 22C3 pharmDx could be defined as moderate to poor For TPS, 22C3 LDT more concordant with the SP263 assay than with the 22C3 pharmDx assay Concordance is slightly increasing at higher cutoffs with CPS SP263 tends to stain more cases than 22C3 	<ul style="list-style-type: none"> No special limitations; the study reports also a good concordance among two scoring pathologists, a higher agreement between TMA and whole sections for TPS than CPS and significant intratumor heterogeneity

(Continues)

TABLE 1 (Continued)

Author, year (Country)	N cases	Sites	Scoring system	Clones	Main findings	Main limitations/Notes
Downes, 2020 (Canada)	27	HNSCC NOS	CPS \geq 1, IC \geq 1%, TC \geq 25%	<ul style="list-style-type: none"> SP142 on Ventana Benchmark Ultra SP263 on Ventana Benchmark Ultra 22C3 pharmDX on a Dako Autostainer Link 48 E1L3N on Leica Bond-III Autostainer 	<ul style="list-style-type: none"> When a lower threshold is used (CPS with 22C3) a higher proportion of cases is positive Interobserver OPA 78–96% and κ 0.517–0.836 higher with SP263, lower with SP142 Intraobserver agreement almost perfect with all assays 	<ul style="list-style-type: none"> No direct comparison among assays All the cutoffs are applied to E1L3N Similar findings also in the cohorts of breast and urothelial cancer
Frederick, 2020 (USA)	NS	HNSCC NOS	CPS 1, 10, 20	22C3 pharmDx	<ul style="list-style-type: none"> Survey of studies on intraobserver and interobserver concordance among 42 pathologists after training in assess CPS 73% and 93% of pathologist met all the endpoint for intra- and interobserver reproducibility 	<ul style="list-style-type: none"> Abstract only Several types of cancer other than HN
Gatalica, 2016 (USA)	NA	HNSCC NOS	5% cut-of	<ul style="list-style-type: none"> SP142, Spring Biosciences SP263, Ventana 22C3, DAKO 28–8, DAKO 	Reported 90% concordance among the 4 assays	<ul style="list-style-type: none"> Abstract only Several cancer types other than HNSCC
Hodgson, 2018 (Canada)	27	HP	Positive if \geq 25% TC with SP263, or if \geq 50% TC or 10% IC with SP142, or if \geq 50% TC with 22C3	<ul style="list-style-type: none"> SP142 on the Ventana Benchmark Ultra SP263 on the Ventana Benchmark Ultra E1L3N on Ventana Benchmark Ultra All three with the Optiview DAB IHC detection kit for all the clones, followed by the Optiview amplification kit for SP142 only 22C3 with the DAKO EnVision FLEX system on a DAKO Autostainer Link 48 system 	<ul style="list-style-type: none"> Substantial agreement between any of the two antibody clones compared with kappa ranging 0.639–0.791 High ICC among assays for TC (\geq0.90), lower for IC (ICC 0.51–0.86) SP142 scoring highest IC, SP263 scoring highest TC 	<ul style="list-style-type: none"> Results for both HN SCC and urothelial cancer together TMA only
Jamshidi, 2020 (USA)	22	HNSCC NOS	CPS 1 and 20	SP263	<ul style="list-style-type: none"> 95% concordance and κ = 0.78 at CPS1 86% concordance and κ = 0.81 at CPS20 Four pathologists scoring 	Abstract only
Koppel, 2018 (Germany)	30	HNSCC and ADK	TC cutoffs 1%, 5%, 10%, 50%	<ul style="list-style-type: none"> Same clone (28–8) with different platforms: <ul style="list-style-type: none"> pharmDx kit Autostainer Link 48 Autostainer Link 48 Dako Omnis Leica Bond-III and Bond Polymer Refine Detection Ventana Benchmark ULTRA with OptiView DAB IHC Detection Kit 	<ul style="list-style-type: none"> 100% inter- and intra-assay repeatability scoring/classification, with acceptable coefficient of variation in percentages of positively stained target cells OPA ranging 83%–100% for HNSCC cases between the various platforms and the reference pharmDx kit Higher concordance for higher cutoffs 	<ul style="list-style-type: none"> Also NSCLC and melanoma cases Not explored interobserver concordance TMA study TC only

(Continues)

TABLE 1 (Continued)

Author, year (Country)	N cases	Sites	Scoring system	Clone	Main findings	Main limitations/Notes
Kringsfeld, 2020 (USA)	305	HNSCC NOS	Cutoffs 1%, 5%, 10%, 25%, 50%	<ul style="list-style-type: none"> 28–8 pharmDx Dako 22C3 pharmDx Dako SP142 Ventana 	Agreement ranging 81%–100% between 28–8 and 22C3 with $\kappa = 0.82$ at cutoff 1%	Also NSCLC, melanoma and urothelial cancer with similar agreement rates and correlation
Nielsen, 2020 (USA)	NA	HNSCC NOS	NA	SP263 on three platforms	98% OPA inter- and intraobserver 99% OPA among platforms	Abstract only; type of cases and specimens not declared
Ramos, 2019 (Colombia)	38	HNSCC NOS	TPS and CPS \geq 1	<ul style="list-style-type: none"> Ventana SP263 Dako 22C3 	Ventana's antibody showed a more intense staining, facilitating the overall assessment of PD-L1 expression	Abstract only, no reported quantitative measures
Ratcliffe, 2016 (UK)	108	HNSCC NOS	1%, 10%, 25%	<ul style="list-style-type: none"> SP263 Ventana 28–8 Dako 22C3 Dako 	<ul style="list-style-type: none"> Correlation coefficient of ≥ 0.9 for each pairwise comparison OPA >90% among the three assays across multiple clinically relevant cut points 	Abstract only
Rebello, 2016 (USA)	100	HNSCC NOS	Positive if $\geq 25\%$ TC	<ul style="list-style-type: none"> SP263 Ventana Benchmark Ultra platform and with the Optiview DAB Detection Kit 	<ul style="list-style-type: none"> OPA 90.8% between three pathologists Intraobserver OPA 94.3% Good analytical properties of the assay in terms of precision, in the range of the 22C3 and 28–8 assays 	Also NSCLC cases, where the cutoff of 25% positive TC with SP263 can discriminate responders to durvalumab
Scott, 2018 (UK)	486	HNSCC NOS	TC, IC and CPS 1 and 10	<ul style="list-style-type: none"> SP263 Ventana SP142 Ventana 22C3 pharmDx 28–8 pharmDx 	<ul style="list-style-type: none"> Good analytical correlation for TC staining for SP263, 22C3 and 28–8 assays SP263 appearing more sensitive, with higher proportion of positive cases OPA between SP263 and other assays at CPS\geq1 ranging 69%–83% 	Abstract only
Vainer, 2019 (Israel)	11	HNSCC NOS	CPS \geq 1	<ul style="list-style-type: none"> 22C3 antibody LDT for the BenchMark XT 22C3 pharmDx 	ICC 0.83 among the two assays	Abstract only
Wang, 2018 (USA)	49	HNSCC HP and SGC	Positive if $\geq 25\%$ TC or IC with SP263, if $\geq 5\%$ IC for SP142	<ul style="list-style-type: none"> SP263 SP142 Both on Benchmark Ultra platform and with the Optiview DAB Detection Kit followed by the Optiview Amplification Kit for SP142 only 	<ul style="list-style-type: none"> Moderate interobserver agreement for both SP263 ($\kappa = 0.469$) and SP142 clones ($\kappa = 0.591$) OPA 80% for SP263 and 88% for SP142 	<ul style="list-style-type: none"> No CPS, cutoffs for urothelial cancer applied to HNSCC No comparison among clones Additional finding of substantial agreement between TMA and WS in ($\kappa = 0.667$)

Abbreviations: ADK, adenocarcinoma; CPS, combined proportion score; HNSCC, head and neck squamous cell carcinoma; HP, hypopharyngeal; IC, immune cells; L, laryngeal; LDT, laboratory developed test; NOS, not otherwise specified; NSCLC, non-small cell lung cancer; OC, oral cavity; OP, oropharynx; OPA, overall percent agreement; SGC, salivary gland cancer; TC, tumor cells; TMA, tissue micro array; TPS, tumor proportion score; WS, whole section.

that concordance between 22C3 pharmDx and SP263 is lower than moderate with ICC and in the range of fair (0.20–0.40) with Cohen's kappa both at cutoff 1 and 20 with no significant increase at the highest cutoff. The LDT of the study, a 22C3 clone on the Ventana Benchmark Ultra platform, showed an ICC of at least moderate in both the comparisons and concordance kappa from fair to substantial with high variability. De Ruiter et al. tested the concordance among the three assays, not only with the clinically relevant cutoffs of CPS1 and 20, but also at the cutoffs of 1% and 50% with TPS, providing an opportunity for comparison with other studies. The concordance of SP263 and 22C3 LDT with 22C3 pharmDx was all in the range of moderate (0.40–0.60), and interestingly, it decreased at a higher cutoff for both the assays against the reference assay.

The impression that evaluation of CPS with SP263 yields a higher quota of positive (CPS \geq 1) cases has also been encountered in recent abstracts,^{22,23} and previous studies focused primarily on concordance among pathologists. Studies arising from a Canadian group of researchers involved a broad population of cancer cases among where 27 HNSCC mixed-site cases were stained with 22C3 pharmDx, SP142 and SP263 and clone E1L3N and evaluated with cutoff CPS \geq 1²⁴ or different percentages of positive immune or tumor cells.^{24–26} These Canadian studies note that when a lower cutoff is applied (as represented by the intrinsic formulation of CPS, where positive immune cells are added to the numerator while the denominator comprises only tumor cells), a greater quota of cases is labeled as positive.²⁴ They also found a high concordance with Cohen's kappa in the range of substantial (0.60–0.80) for all pairwise comparisons and excellent ICC (higher than 0.90) for all clones with exception of SP142 when specifically considering concordance on tumor cell staining.²⁵ This is, however, in line with findings in other cancer sites where SP142 is known to better stain immune cells than tumor cells.²⁷ Studies from De Meulenaere et al., where the SP142 assay is used parallel to the reference 22C3 even if with no direct comparison, a tendency emerged toward lower rate of positivity in tumor cells with the SP142.^{28,29}

Few studies and published abstracts have investigated the assay 28–8 run on the Dako platform, in comparison with other available assays.^{30–33} Even though only overall percent agreement (OPA) was reported, these studies suggest a high degree of concordance among 28–8 and other assays, with OPA ranging from 80% to 100%,^{30–33} correlation coefficient higher than 0.9,³² and kappa of 0.82 which is considered almost perfect.³¹ The 28–8 clone was also tested on platforms different from the Dako Autostainer Link 48 platform approved as a complementary diagnostic.³⁴ In this study, the platforms used were Dako Autostainer Link 48, Dako Omnis, Leica Bond-III, and Ventana Benchmark Ultra. The authors found an acceptable coefficient of variation in the percentages of positively stained target cells, with OPA ranging from 83% to 100% specifically for HNSCC cases between the various platforms and the reference pharmDx kit, and increasing concordance for higher cutoffs.³⁴ Studies such as the one reported by Koppel et al. shed light on common problems in harmonization and comparability studies, because of the multiplicity of platforms deployed and staining protocols applied, especially when

multicenter comparison studies are attempted. This yields additional uncertainty, including the critical evaluation of published evidence, as no true direct comparison among results from studies is possible.

In summary, the main published findings to date support a variable degree of concordance from fair to substantial among the reference assay 22C3 pharmDx and two alternate Ventana assays, with the caveat that SP263 is likely to label more cases as positive due to stronger and more diffuse staining of tumor cells, while SP142 stains immune cells more so than tumor cells which in some cases can result in a lower CPS. While limited evidence is available concerning the 28–8 assay, available findings show very good concordance with this assay and all others so far tested.

3.2 | Concordance among pathologists

Another important point when assessing PD-L1 expression, especially with the formulation of CPS, is concordance among pathologists. Indeed, CPS is more complex and perhaps less intuitive than TPS, as it requires specific counting of tumor and immune cells to calculate the score. Not surprisingly, training in this regard has been shown to be important.³⁵ However, when pathologists are trained, the reproducibility among them appears to be high in assessing CPS, with ICC \geq 0.70^{21,29} or excellent \geq 0.90²⁰ with all of the assays including 22C3 pharmDx, SP263, and SP142. Some studies suggest better concordance for pathologists with the SP263 assay when assessing both CPS or separately counting tumor and immune cells, with an almost perfect kappa of 0.836 and in the range of moderate to substantial for the clones SP142, 22C3, and E1L3N.²⁴ Similar results with very high concordance among scoring pathologists with SP263 assay have recently appeared in some abstracts.^{36,37} Ease of use of the SP263 assay was previously recognized in validation studies. For example, Rebelatto et al. demonstrated the ease of use when the SP263 assay was validated in a mixed cohort of lung cancer and HNSCC cases.³⁸ The assay not only showed good analytical performance in terms of precision and robustness against preanalytical factors, but also very high interobserver and intraobserver concordance. The authors stated that the SP263 assay has analytical and reading properties in the same range of the two approved 22C3 and 28–8 assays, thus meeting the requirement criteria for assay use. Important to note, while the approved assays for HNSCC 22C3 and 28–8 use a clone that detects the epitopes of PD-L1 located in the extracellular domain, the Ventana assays and the clone E1L3N detect epitopes within the intracellular domain. It has been recognized that antibodies raised against the cytoplasmic domain of PD-L1 offer better visualization of membrane PD-L1 compared with those raised against the extracellular domain.³⁹ This may explain, at least partially, the difference in staining pattern and the preferential staining of tumor cells with SP263.

Finally, an important aid for pathologists to interpret PD-L1 expression could come from artificial intelligence (AI). As stated in a recent review by Inge et al.⁴⁰, AI algorithms are already being used to investigate the role of PD-L1, providing considerable insight into its

expression and heterogeneity within the tumor microenvironment. Even though the few studies to date which have investigated AI algorithms for PD-L1 scoring dealt with cancer types other than HNSCC, the findings are interesting when bearing in mind that automated algorithms could be trained to recognize different types of cells (tumor vs. immune; with macrophages to be excluded from other cell types) and provide a more objective quantification of CPS. In an AI study assessing CPS with image analysis in gastric cancer, the authors showed that application of their IHC membrane algorithm for PD-L1 evaluation was 85% concordant with manual scoring.⁴¹ Furthermore, PD-L1 AI scores were comparable with manual scoring in predicting patient response to pembrolizumab.⁴¹ The challenges encountered by the algorithms in recognizing and discriminating immune cells to be counted are the same as those encountered by pathologists. Hence, caution is warranted. It is foreseeable that, as transpired with using image analysis to assess breast biomarkers,⁴² the incorporation of clinical outcome as one of the benchmarks for algorithms could enable development of more clinically relevant tools.

4 | CONCLUSION

The approval of ICI for HNSCC triggered by positive expression of PD-L1, and concurrent availability of several assays with different characteristics, prompted researchers to undertake comparative studies and explore interchangeability of these assays. While several such studies have been published in the last five years, only a small proportion dealt with HNSCC cases and incorporated varying cutoffs for positivity with or without evaluating immune and tumor cells. Only in the last year have studies appeared with a research design clearly focused on comparing of two or more assays with the reference 22C3 assay, and employing the required CPS scoring system. Emerging results currently suggest a moderate degree of concordance between the reference standard and widely used SP263 Ventana assay. Moreover, differences in staining patterns among these assays are observed, with SP263 demonstrating more strong staining of tumor cells and SP142 better staining of immune cells. However, heterogeneity in these varied studies (e.g., preanalytical factors, platforms used, detection systems applied, one vs multiple pathologist assessment) hamper critical and robust comparison of their results. Clearly, further work is needed to draw stronger conclusions on the interchangeability of PD-L1 assays in HNSCC.

CONFLICT OF INTEREST

The authors declare that they do not have any conflict of interest.

AUTHOR CONTRIBUTIONS

Ilaria Girolami: Conceptualization; Data curation; Formal analysis; Methodology; Writing-original draft. Liron Pantanowitz: Conceptualization; Investigation; Methodology; Writing-original draft; Writing-review & editing. Massimo Barberis: Conceptualization; Visualization; Writing-review & editing. Gaetano

Paolino: Data curation; Visualization; Writing-review & editing. Matteo Brunelli: Conceptualization; Visualization; Writing-review & editing. Elena Vigliar: Visualization; Writing-review & editing. Enrico Munari: Visualization; Writing-review & editing. Swati Satturwar: Visualization; Writing-review & editing. Giancarlo Troncone: Conceptualization; Visualization; Writing-review & editing. Albino Eccher: Conceptualization; Data curation; Methodology; Project administration; Resources; Supervision; Writing-original draft.

ETHICAL APPROVAL

No institutional review board approval was needed, as no ethical issue is raised by literature reviews.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/jop.13220>.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

ORCID

Albino Eccher  <https://orcid.org/0000-0002-9992-5550>

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-249. <https://doi.org/10.3322/caac.21660>
2. Pulte D, Brenner H. Changes in survival in head and neck cancers in the late 20th and early 21st century: a period analysis. *Oncologist*. 2010;15:994-1001. <https://doi.org/10.1634/theoncologist.2009-0289>
3. Johnson DE, Burtneß B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. *Nat Rev Dis Prim*. 2020;6:92. <https://doi.org/10.1038/s41572-020-00224-3>
4. Cohen EEV, Soulières D, Le Tourneau C, et al. Pembrolizumab versus methotrexate, docetaxel, or cetuximab for recurrent or metastatic head-and-neck squamous cell carcinoma (KEYNOTE-040): a randomised, open-label, phase 3 study. *Lancet*. 2019;393:156-167. [https://doi.org/10.1016/S0140-6736\(18\)31999-8](https://doi.org/10.1016/S0140-6736(18)31999-8)
5. Burtneß B, Harrington KJ, Greil R, et al. Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): a randomised, open-label, phase 3 study. *Lancet*. 2019;394:1915-1928. [https://doi.org/10.1016/S0140-6736\(19\)32591-7](https://doi.org/10.1016/S0140-6736(19)32591-7)
6. Canning M, Guo G, Yu M, et al. Heterogeneity of the head and neck squamous cell carcinoma immune landscape and its impact on immunotherapy. *Front Cell Dev Biol*. 2019;7:52. <https://doi.org/10.3389/fcell.2019.00052>
7. Ferris RL, Blumenschein GJ, Fayette J, et al. Nivolumab vs investigator's choice in recurrent or metastatic squamous cell carcinoma of the head and neck: 2-year long-term survival update of CheckMate 141 with analyses by tumor PD-L1 expression. *Oral Oncol*. 2018;81:45-51. <https://doi.org/10.1016/j.oraloncology.2018.04.008>
8. Harrington KJ, Ferris RL, Blumenschein G, et al. Nivolumab versus standard, single-agent therapy of investigator's choice in recurrent

- or metastatic squamous cell carcinoma of the head and neck (CheckMate 141): health-related quality-of-life results from a randomised, phase 3 trial. *Lancet Oncol*. 2017;18:1104-1115. [https://doi.org/10.1016/S1470-2045\(17\)30421-7](https://doi.org/10.1016/S1470-2045(17)30421-7)
9. Ferris RL, Blumenschein GJ, Fayette J, et al. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *N Engl J Med*. 2016;375:1856-1867. <https://doi.org/10.1056/NEJMoa1602252>
 10. Ferris RL, Licitra L, Fayette J, et al. Nivolumab in patients with recurrent or metastatic squamous cell carcinoma of the head and neck: efficacy and safety in checkmate 141 by prior cetuximab use. *Clin Cancer Res*. 2019;25:5221-5230. <https://doi.org/10.1158/1078-0432.CCR-18-3944>
 11. Ries J, Agaimy A, Wehrhan F, et al. Importance of the PD-1/PD-L1 axis for malignant transformation and risk assessment of oral leukoplakia. *Biomedicine*. 2021;9:194. <https://doi.org/10.3390/biomedicines9020194>
 12. Girolami I, Pantanowitz L, Munari E, et al. Prevalence of PD-L1 expression in head and neck squamous precancerous lesions: a systematic review and meta-analysis. *Head Neck*. 2020;42:3018-3030. Published online June 22, 2020;hed.26339. <https://doi.org/10.1002/hed.26339>
 13. Cramer JD, Burtneß B, Ferris RL. Immunotherapy for head and neck cancer: recent advances and future directions. *Oral Oncol*. 2019;99:104460-<https://doi.org/10.1016/j.oraloncology.2019.104460>
 14. Emancipator K, Huang L, Aurora-Garg D, et al. Comparing programmed death ligand 1 scores for predicting pembrolizumab efficacy in head and neck cancer. *Mod Pathol*. 2021;34:532-541. <https://doi.org/10.1038/s41379-020-00710-9>
 15. Concha-Benavente F, Srivastava RM, Trivedi S, et al. Identification of the cell-intrinsic and -extrinsic pathways downstream of EGFR and IFN γ that induce PD-L1 expression in head and neck cancer. *Cancer Res*. 2016;76:1031-1043. <https://doi.org/10.1158/0008-5472.CAN-15-2001>
 16. Ancevski Hunter K, Socinski MA, Villaruz LC. PD-L1 testing in guiding patient selection for pd-1/pd-l1 inhibitor therapy in lung cancer. *Mol Diagn Ther*. 2018;22:1-10. <https://doi.org/10.1007/s40291-017-0308-6>
 17. Food and Drug Administration. PD-L1 IHC 22C3 pharmDx interpretation manual - Head and Neck Squamous Cell Carcinoma (HNSCC). https://www.agilent.com/cs/library/usermanuals/public/29314_22c3_pharmDx_hnsccl_interpretation_manual_us.pdf. Accessed March 24, 2021.
 18. Food and Drug Administration. PD-L1 IHC 28-8 pharmDx interpretation manual - Head and Neck Squamous Cell Carcinoma (HNSCC). https://www.agilent.com/cs/library/usermanuals/public/29186_pd-l1-ihc-28-8-interpretation-manual-scchn.pdf. Accessed March 24, 2021.
 19. Ionescu DN, Downes MR, Christofides A, Tsao MS. Harmonization of PD-L1 testing in oncology: a Canadian pathology perspective. *Curr Oncol*. 2018;25:e209-e216. <https://doi.org/10.3747/co.25.4031>
 20. de Ruiter EJ, Mulder FJ, Koomen BM, et al. Comparison of three PD-L1 immunohistochemical assays in head and neck squamous cell carcinoma (HNSCC). *Mod Pathol*. 2021;34:1125-1132. <https://doi.org/10.1038/s41379-020-0644-7>
 21. Crosta S, Boldorini R, Bono F, et al. PD-L1 testing and squamous cell carcinoma of the head and neck: a multicenter study on the diagnostic reproducibility of different protocols. *Cancers (Basel)*. 2021;13:292. <https://doi.org/10.3390/cancers13020292>
 22. Ramos M, Baldion A, Suarez D, et al. Immunohistochemistry analysis of PD-L1 expression in head and neck cancer. *Virchows Arch*. 2019;475(S1):1-436. <https://doi.org/10.1007/s00428-019-02631-8>
 23. Al-Masri H, Ratcliffe M, Sharpe A, et al. Concordance of tumour and immune cell staining with Ventana SP142, Ventana SP263, Dako 22C3 and Dako 28-8 PD-L1 tests across different cancer types. *Virchows Arch*. 2017;471(S1):1-352. <https://doi.org/10.1007/s00428-017-2205-0>
 24. Downes MR, Slodkowska E, Katabi N, Jungbluth AA, Xu B. Inter- and intraobserver agreement of programmed death ligand 1 scoring in head and neck squamous cell carcinoma, urothelial carcinoma and breast carcinoma. *Histopathology*. 2020;76:191-200. <https://doi.org/10.1111/his.13946>
 25. Hodgson A, Slodkowska E, Jungbluth A, et al. PD-L1 immunohistochemistry assay concordance in urothelial carcinoma of the bladder and hypopharyngeal squamous cell carcinoma. *Am J Surg Pathol*. 2018;42:1059-1066. <https://doi.org/10.1097/PAS.0000000000001084>
 26. Wang C, Hahn E, Slodkowska E, et al. Reproducibility of PD-L1 immunohistochemistry interpretation across various types of genitourinary and head/neck carcinomas, antibody clones, and tissue types. *Hum Pathol*. 2018;82:131-139. <https://doi.org/10.1016/j.humpath.2018.07.024>
 27. 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;12:208-222. <https://doi.org/10.1016/j.jtho.2016.11.2228>
 28. De Meulenaere A, Vermassen T, Aspeslagh S, et al. Tumor PD-L1 status and CD8(+) tumor-infiltrating T cells: markers of improved prognosis in oropharyngeal cancer. *Oncotarget*. 2017;8:80443-80452. <https://doi.org/10.18632/oncotarget.19045>
 29. De Meulenaere A, Vermassen T, Creyten D, et al. Importance of choice of materials and methods in PD-L1 and TIL assessment in oropharyngeal squamous cell carcinoma. *Histopathology*. 2018;73:500-509. <https://doi.org/10.1111/his.13650>
 30. Gatalica Z, Vanderwalde AM, Rose I, et al. Distribution of PD-L1 expression in diverse cancer types: experience with over 10,000 cases. *J Clin Oncol*. 2016;34(15_suppl):11548. https://doi.org/10.1200/JCO.2016.34.15_suppl.11548
 31. Krigsfeld GS, Prince EA, Pratt J, et al. Analysis of real-world PD-L1 IHC 28-8 and 22C3 pharmDx assay utilisation, turnaround times and analytical concordance across multiple tumour types. *J Clin Pathol*. 2020;73:656-664. <https://doi.org/10.1136/jclinpath-2020-206466>
 32. Ratcliffe M, Sharpe A, Rebelatto M, et al. A comparative study of PD-L1 diagnostic assays in squamous cell carcinoma of the head and neck (SCCHN). *Ann Oncol*. 2016;27:vi328-vi350. <https://doi.org/10.1093/annonc/mdw376.7>
 33. Scott M, Wildsmith S, Ratcliffe M, et al. Comparison of patient populations identified by different PD-L1 assays in head and neck squamous cell carcinoma (HNSCC). *Ann Oncol*. 2018;29:375.
 34. Koppel C, Schwellenbach H, Zielinski D, et al. Optimization and validation of PD-L1 immunohistochemistry staining protocols using the antibody clone 28-8 on different staining platforms. *Mod Pathol*. 2018;31:1630-1644. <https://doi.org/10.1038/s41379-018-0071-1>
 35. Eccher A, Fontanini G, Fusco N, et al. Digital slides as an effective tool for programmed death ligand 1 combined positive score assessment and training: Lessons learned from the "Programmed death ligand 1 key learning program in Head-and-Neck squamous cell carcinoma". *J Pathol Inform*. 2021;12:1. https://doi.org/10.4103/jpi.jpi_63_20
 36. Jamshidi P, Kaur A, Liu L, Sullivan M, Watkin W, Paintal A. Inter-Rater Reliability (IRR) in a Consensus PD-L1 Immunohistochemistry (IHC) service in an academic multi-hospital health system. *Mod Pathol*. 2020;33(suppl 2):1863.
 37. Nielsen A, Manriquez G, Hayden D, et al. Precision and repeatability of the VENTANA PD-L1 (SP263) assay across six different tumor types. *Mod Pathol*. 2020;33(suppl 2):830.
 38. Rebelatto MC, Midha A, Mistry A, et al. Development of a programmed cell death ligand-1 immunohistochemical assay validated for analysis of non-small cell lung cancer and head and neck

- squamous cell carcinoma. *Diagn Pathol.* 2016;11:95. <https://doi.org/10.1186/s13000-016-0545-8>
39. Mahoney KM, Sun H, Liao X, et al. PD-L1 antibodies to its cytoplasmic domain most clearly delineate cell membranes in immunohistochemical staining of tumor cells. *Cancer Immunol Res.* 2015;3:1308-1315. <https://doi.org/10.1158/2326-6066.CIR-15-0116>
40. Inge LJ, Dennis E. Development and applications of computer image analysis algorithms for scoring of PD-L1 immunohistochemistry. *Immuno-Oncology Technol.* 2020;6:2-8. <https://doi.org/10.1016/j.iotech.2020.04.001>
41. Kim H-N, Jang J, Heo YJ, et al. PD-L1 expression in gastric cancer determined by digital image analyses: pitfalls and correlation with pathologist interpretation. *Virchows Arch.* 2020;476:243-250. <https://doi.org/10.1007/s00428-019-02653-2>
42. Barnes M, Srinivas C, Bai I, et al. Whole tumor section quantitative image analysis maximizes between-pathologists' reproducibility

for clinical immunohistochemistry-based biomarkers. *Lab Invest.* 2017;97:1508-1515. <https://doi.org/10.1038/labinvest.2017.82>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Girolami I, Pantanowitz L, Barberis M, et al. Challenges facing pathologists evaluating PD-L1 in head & neck squamous cell carcinoma. *J Oral Pathol Med.* 2021;50:864-873. <https://doi.org/10.1111/jop.13220>