Histopathology

Histopathology 2020, **76,** 793–802. DOI: 10.1111/his.14040

REVIEW

Comparability of PD-L1 immunohistochemistry assays for non-small-cell lung cancer: a systematic review

Bregje M Koomen,[1] Sushil K Badrising,[2] Michel M van den Heuvel[2] & Stefan M Willems[1]

[1]*Department of Pathology, University Medical Center Utrecht, Utrecht, and* [2]*Department of Pulmonary Diseases, Radboudumc, Nijmegen, the Netherlands*

Koomen B M, Badrising S K, van den Heuvel M M & Willems S M.

(2020) *Histopathology* **76**, 793–802. https://doi.org/10.1111/his.14040

Comparability of PD-L1 immunohistochemistry assays for non-small-cell lung cancer: a systematic review

Programmed cell death ligand 1 (PD-L1) immunohistochemistry is used to determine which patients with advanced non-small-cell lung cancer (NSCLC) respond best to treatment with PD-L1 inhibitors. For each inhibitor, a unique immunohistochemical assay was developed. This systematic review gives an up-to-date insight into the comparability of standardised immunohistochemical assays and laboratory-developed tests (LDTs), focusing specifically on tumour cell (TC) staining and scoring. A systematic search was performed identifying publications that assessed interassay, interobserver and/or interlaboratory concordance of PD-L1 assays and LDTs in tissue of NSCLC patients. Of 4294 publications identified through the systematic search, 27 fulfilled the inclusion criteria and were of sufficient methodological quality. Studies assessing interassay concordance found high agreement between assays 22C3, 28-8 and SP263 and properly validated LDTs, and lower concordance for comparisons involving SP142. A decrease in concordance, however, is seen with use of cut-offs, which hampers interchangeability of PD-L1 immunohistochemistry assays and LDTs. Studies assessing interobserver concordance found high agreement for all assays and LDTs, but lower agreement with use of a 1% cut-off. This may be problematic in clinical practice, as discordance between pathologists at this cut-off may result in some patients being denied valuable treatment options. Finally, five studies assessed interlaboratory concordance and found moderate to high agreement levels for various assays and LDTs. However, to assess the actual existence of interlaboratory variation in PD-L1 testing and PD-L1 positivity in clinical practice, studies using real-world clinical pathology data are needed.

Keywords: immunohistochemistry, immunotherapy, non-small-cell lung cancer, predictive biomarker, programmed cell death-ligand 1, systematic review

Introduction

Since the approval of the first immune check-point inhibitor in 2011,[1-3] immunotherapy has become an important part of treatment for several forms of cancer. In patients with advanced non-small-cell lung cancer (NSCLC), treatment with programmed cell death-1 (PD-1) or programmed cell death-ligand 1 (PD-L1) inhibitors has become part of standard care. These patients may be treated with nivolumab or pembrolizumab, both anti-PD-1 check-point inhibitors, or with an anti-PD-L1 check-point inhibitor, i.e. atezolizumab or durvalumab.[4-9] Some of these drugs may only be prescribed to patients who show PD-L1 expression in at least 1% or 50% of tumour cells,

Address for correspondence: B M Koomen, In-hospital mail: H04.312, Department of Pathology, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands. e-mail: b.m.koomen@umcutrecht.nl

measured with immunohistochemistry (IHC).[10-12] Immunohistochemical PD-L1 testing thus aids clinicians in treatment decision-making.

For each immune check-point inhibitor, however, a separate immunohistochemistry (IHC) PD-L1 assay has been developed. The PD-L1 IHC 22C3 PharmDx assay was used in clinical trials assessing efficacy of pembrolizumab, and is therefore Food and Drug Administration (FDA)-approved and Conformité Européenne (CE)-marked as a companion diagnostic for prescription of this drug.[8,13,14] In a similar fashion, the PD-L1 IHC 28-8 PharmDx assay was FDA-approved and CE-marked as a complementary diagnostic for nivolumab,[15,16] while the PD-L1 IHC SP142 assay became a complementary diagnostic for atezolizumab.[17,18] Finally, the PD-L1 IHC SP263 assay was developed for durvalumab, but it has also received CE marking for identification of patients eligible for treatment with pembrolizumab and of patients most likely to benefit from treatment with nivolumab.[19,20]

Using all these different assays to test for PD-L1 expression in one pathology laboratory is not feasible. Not only would it be expensive and time-consuming to run so many different tests for each patient, most laboratories will not have both staining platforms (i.e. Dako and Ventana/Roche) needed for these tests at their disposal. Furthermore, the number of tests that can be performed is restricted due to limited tissue availability.[21] It is thus important to assess whether results from different assays are interchangeable. In addition, it should be assessed if laboratory-developed tests (LDTs) can be used instead of the standardised PD-L1 assays. In recent years, a multitude of studies examining these issues has been published, such as the Blueprint PD-L1 IHC Assay Comparison Project[22] or the harmonisation studies by Ratcliffe *et al.*,[23] Rimm *et al.*[24] and Scheel *et al.*[25] Others, such as Büttner *et al.*,[19] have reviewed the analytical performance of PD-L1 IHC assays previously. Considering the abundance of studies that have been published on the subject, however, there is need for a systematic, comprehensive and up-to-date overview of the literature, which not only focuses on interassay and interobserver concordance, but also includes a review of interlaboratory concordance. Hence, the aim of this study was to systematically review all studies that assessed interassay, interobserver and/or interlaboratory concordance of PD-L1 IHC assays and LDTs, and in so doing provide an updated insight into the comparability of these standardised assays and LDTs.

Materials and methods

SEARCH STRATEGY

A systematic search of PubMed, Embase and Cochrane Library was performed, using the search terms ‘lung cancer’ and ‘PD-L1’ with all relevant synonyms (see Table S1). Only these two terms were used to ensure that no relevant articles would be missed. Adding another term, such as ‘immunohistochemistry’, might have made the search more specific, but would also have increased the risk of eliminating relevant titles. After removal of duplicates, titles and abstracts were screened by two researchers independently (B.K. and S.B.) based on predefined inclusion and exclusion criteria (see Table S2). Remaining articles were read in full, and a further selection was made based on the relevance of these full texts. Discrepancies between the two researchers were discussed and resolved by consensus.

INCLUSION CRITERIA

Studies were included if they evaluated interassay, interobserver and/or interlaboratory concordance of at least two PD-L1 IHC assays and/or LDTs used on tissue from NSCLC patients in clinical practice. Studies examining interobserver and/or interlaboratory concordance in only one assay were also included. In order for studies to qualify, determination of PD-L1 expression had to be performed on histological tissue from NSCLC patients and appropriate scoring methods had to be used (i.e. assessment of membranous staining of tumour cells by at least one pathologist). Since PD-L1 IHC was validated in histological specimens, studies examining cytological specimens only were excluded. Studies that did not perform adequate statistical analysis to compare assays (i.e. overall percentage of agreement should at least be given) were also excluded. Only articles written in English and containing original published data were eligible for inclusion.

QUALITY ASSESSMENT

Methodological quality of all articles remaining after full text reading were appraised by using a revised form of the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) tool for assessing risk of bias.[26] Originally, this tool consists of four domains, i.e. patient selection, index test, reference standard and flow and timing. As individual PD-L1 IHC assays were not compared to a reference standard in the included studies, but rather to each other, the reference standard domain was excluded from the QUADAS-2 tool for this review. Instead, another domain was added based on the Quality in Prognosis Studies (QUIPS) tool, i.e. statistical analysis and reporting.[27] Risk of bias was scored as low, moderate or high for each domain of the revised QUADAS-2 tool and points were awarded accordingly (1 point for low risk of bias, 0.5 points for moderate risk of bias and 0 points for high risk of bias). Based on the sum of the scores given to each individual domain, overall scores of low, moderate or high risk of bias were awarded to studies using the following scoring system: low risk of bias for studies with ≥3.5 points, moderate risk of bias for studies with ≥2.5 and <3.5 points and high risk of bias for studies with <2.5 points. Appraisal of methodological quality was performed independently by two researchers (B.K. and S.B.) and differences were resolved through discussion. Studies with high risk of bias were excluded from data extraction and further analysis.

DATA EXTRACTION AND SYNTHESIS OF RESULTS

The following data were extracted from each study included after appraisal of methodological quality: first author's name, year of publication, sample size, type of cancer of included patients, type of material used for PD-L1 testing, type of standardised assay and/or LDT used for PD-L1 testing, scoring method, cut-off values, number of observers scoring PD-L1, type of statistical analysis and results from comparison between assays, observers and/or laboratories. This review focuses on concordance of tumour cell (TC) staining and scoring, as treatment decisions for NSCLC patients are based on scoring of PD-L1 expression on TCs in clinical practice. However, as scoring of PD-L1 expression on immune cells (IC) could become relevant to clinical practice in the future, we also extracted data on concordance of IC staining and scoring and included this as Data S1 and Table S8. Due to heterogeneity between included studies, such as differences in antibodies tested, number of pathologists scoring and statistical methods applied, results could not be quantitatively pooled and a meta-analysis could not be performed.

Results

SYSTEMATIC SEARCH AND STUDY SELECTION

The search in PubMed, Embase and Cochrane Library yielded 4294 unique hits after removal of duplicates (see Figure 1). Fifty-nine records remained after screening of titles and abstracts. Of these, one full text was unavailable. Therefore, 58 full text articles were evaluated in detail, of which 41 articles met the inclusion criteria. All selected articles studied interassay, interobserver and/or interlaboratory concordance of at least one PD-L1 IHC assay, using material from NSCLC patients. Most studies included multiple subtypes of NSCLC, with adenocarcinoma and squamous cell carcinoma being studied most frequently. Some studies also included patients with other types of lung cancer, such as small cell lung cancer (SCLC)[28,29] and mesothelioma.[30] Sample sizes ranged from 15 to 713 tissue specimens. All studies used statistical analysis to measure concordance. The statistical methods used, however, varied. The kappa statistic (κ) was used most, but some studies used intraclass correlation coefficient (ICC), Pearson's/Spearman's correlation or calculation of percentage agreement.

QUALITY ASSESSMENT

The 41 articles selected through full text reading were critically appraised on methodological quality. Based on scoring with the revised QUADAS-2 tool, studies ranged from low to high risk of bias (see Table S3). Studies with high risk of bias were often unclear concerning their method of patient selection and reasons for patient exclusion, about blinding of pathologists for each other's results and for the specific antibody used, about the use of staining platform and staining protocol, or about the scoring method used. Also, some studies did not provide sufficient information on the use of statistical methods or did not present all data, prohibiting assessment of adequacy of analytical strategy. Five studies were judged as having low risk of bias,29,31-34 22 studies as having moderate risk of bias[22-25,28,30,35-50] and 14 studies as having high risk of bias.[51-64] The 14 studies with high risk of bias were excluded, which left 27 articles for data extraction and further analysis. An overview of study characteristics of all included studies can be found in Table S4.

INTERASSAY CONCORDANCE

Of the 27 included articles, 22 reported on interassay concordance of TC staining between PD-L1 IHC assays. A summary of results from all 22 studies can be found in Table 1, while a more detailed presentation of results from each study can be found in Table S5. Many studies compared the standardised assays 22C3, 28-8, SP263 and SP142. Overall,

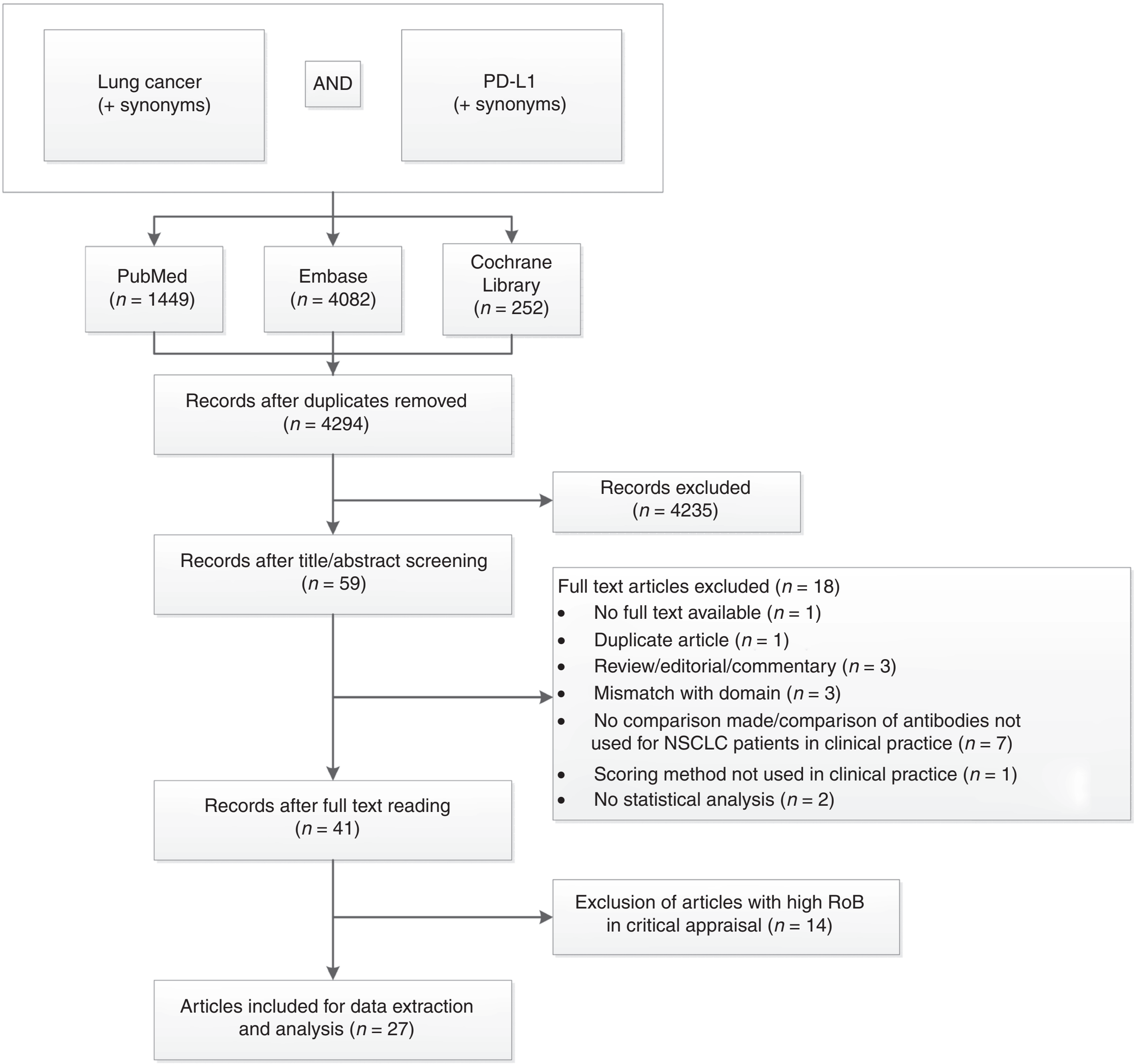
Lung cancer
(+ synonyms)
AND
PD-L1
(+ synonyms)
PubMed
(n = 1449)
Embase
(n = 4082)
Cochrane
Library
(n = 252)
Records after duplicates removed
(n = 4294)
Records excluded
(n = 4235)
Records after title/abstract screening
(n = 59)
Full text articles excluded (n = 18)
• No full text available (n = 1)
• Duplicate article (n = 1)
• Review/editorial/commentary (n = 3)
• Mismatch with domain (n = 3)
• No comparison made/comparison of antibodies not used for NSCLC patients in clinical practice (n = 7)
• Scoring method not used in clinical practice (n = 1)
• No statistical analysis (n = 2)
Records after full text reading
(n = 41)
Exclusion of articles with high RoB
in critical appraisal (n = 14)
Articles included for data extraction
and analysis (n = 27)

Figure 1. Flowchart of study selection process (date of search: 27 June 2018). PD-L1, programmed cell death-ligand 1; RoB, risk of bias.

moderate to strong concordance was seen between 22C3, 28-8 and SP263[22-23,28,31-32,35,38,43,48] and lower concordance between SP142 and the other assays.[22,24,28,31-32,35,38-39,48,50] Concordance was often highest between assays 22C3 and 28-8,[22,30,32,35,48] such as demonstrated by Brunnström *et al.*,[32] who found a weighted κ value of 0.891 (0.82–0.96) for comparison of these two assays. Two studies by Scheel *et al.*[25,46] showed somewhat lower concordance values between 22C3, 28-8 and SP263 than the other studies, but these results may have been affected by interobserver variation and by the low sample size in both studies. Several studies described a higher proportion of stained TCs with use of antibody SP263 when compared to antibody 22C3 and/or 28-8.[25,28,35,44,46,48] According to Munari *et al.*,[44] this difference in staining led to a significantly lower proportion of positive cases with assay 22C3 compared to assay SP263 for both the 1 and 50% cut-offs. Similarly, other studies also assessed concordance with deployment of clinically relevant cut-offs. Some of these studies showed diminished concordance rates when cut-offs were used. Hendry *et al.*[28] showed only moderate agreement between

Table 1. Summary of results from studies assessing interassay concordance of TC staining

Type of test	Comparison	Interassay concordance
Standardised assays	22C3, 28-8 and SP263	• Moderate to high concordance for all comparisons[22-23,28,31-32,35,38,43,48] • Highest concordance between 22C3 and 28-8[22,30,32,35,48] • Lower concordance rates with use of cut-offs[22,28,44], especially using the 1% cut-off[35,38,43]
	SP142 versus all other assays	• Lower concordance levels compared to comparisons between all other assays[22,24,28,31-32,35,38-39,48,50]
LDTs	Various LDTs versus standardised assays	• High concordance for some LDTs, only if appropriate protocol used[31-32,47]
	22C3 LDT versus 22C3 standardised assay	• High correlation[28,40-41,44] • In some studies higher correlation than between two different standardised assays[28,44]
	E1L3N versus all standardised assays	• High concordance between E1L3N and 22C3, 28-8 and SP263[24,37,47] • Lower concordance between E1L3N and SP142[24,42]

LDT, Laboratory-developed test; TC, Tumour cell.

22C3, 28-8 and SP263 when cut-offs were used (Cohen's κ range = 0.433–0.631), while good agreement was found for PD-L1 expression on a continuous scale (ICC range = 0.726–0.812). In the Blueprint Phase 1 study, agreement with the reference assay ranged from 86.8% to 94.7% for comparisons between antibody 22C3, 28-8 and SP263 when different cut-offs were used, meaning that in some cases almost 15% of patients in the study would not have been assigned a treatment if an alternative to the reference assay had been used.[22] Other studies showed lower agreement for the 1% than for the 50% cut-off.[35,38,43] Two studies showed good concordance between assays for any cut-off used,[23,30] but these studies only calculated percentage agreement, which may overestimate true agreement.[65,66]

Comparisons of TC staining were also made between the 22C3, 28-8, SP263 and SP142 antibodies being used with their standard protocols and being used in LDTs. A study by Adam *et al.*[31] demonstrated that 14 of 27 LDTs were concordant (defined as weighted κ value ≥0.75) with one of the pre-specified reference assays. The lowest κ value was seen for the SP142 LDT compared to the SP263 reference assay (weighted κ = 0.38). Two studies by Ilie *et al.*[40,41] showed high correlation between two different 22C3 LDTs and the 22C3 standardised assay. Another study also showed excellent agreement between the 22C3 standardised assay and LDT, with an ICC of 0.921 and Cohen's κ of 0.897 for the 50% cut-off.[28] In this study, discrepancies were actually much greater between two different antibodies used on the same platform (22C3 and 28-8) than between the same antibody (22C3) used on different platforms. A similar finding was reported by Munari *et al.*[44] Other studies also compared one or more of the aforementioned standardised assays with antibody E1L3N, which is used as an LDT by some laboratories in clinical practice. Good correlation was seen between E1L3N and assays SP263, 28-8 and 22C3,[24,37,47] while comparison with SP142 again showed lower concordance values.[24,42] One study[36] showed higher sensitivity in staining of PD-L1 using 28-8 compared to E1L3N. Finally, in a study by Soo *et al.*,[47] which used the SP142 antibody as an LDT, changes in the SP142 protocol led to a higher intensity of staining compared to the original protocol, demonstrating how the IHC protocol can influence the apparent level of PD-L1 expression.

INTEROBSERVER CONCORDANCE

Sixteen of the 27 included studies examined interobserver concordance (see Table 2; Table S6). All these studies assessed concordance between pathologists scoring TC staining and most found moderate to almost perfect agreement for all assays.[23-24,29,32-35,37,39-40,45,48,49] Only Scheel *et al.*[25] found somewhat lower concordance values for E1L3N and SP142 LDTs and 22C3, 28-8, SP263 and SP142 standardised assays when a scoring system applying five cut-offs was used (Light's κ range = 0.47–0.50). However, the sample size in this study was very small (n = 15), and classifying the cases by the dichotomous cut-off criteria included in the scoring system resulted in higher concordance levels for all antibodies (Light's κ range = 0.59–0.80). Other studies also assessed interobserver concordance for multiple cut-

Table 2. Summary of results from studies assessing interobserver concordance of TC scoring

Type of test	Overall	Use of cut-offs
Standardised assays	• Good concordance for all standardised assays[23-24,29,32-35,37,39-40,43-45,48,49] • One study showing only moderate agreement[25]	• Lower concordance levels for 1% cut-off compared to 50% cut-off23-24,29,32,35,43,48 • Lower concordance levels for 1% cut-off compared to 5%, 10% and 25% cut-offs23,32,48 • Lower concordance levels for 80% cut-off compared to other cut-offs48
LDTs	Good concordance for various LDTs24,29,32,37,40,44,45	Lower concordance levels for 1% cut-off compared to other cut-offs24,29,32

LDT, Laboratory-developed test; TC, Tumour cell.

offs, and many found concordance levels to be lower for the 1% cut-off compared to the 50% cut-off[23-24,29,32,35,43,48] and the 5%, 10% or 25% cut-off.[23,32,48] The Blueprint Phase 2 study also assessed the 80% cut-off and found interpathologist agreement to be slightly diminished for this cut-off compared to the 5%, 10%, 25% and 50% cut-offs.[48] A study by Cooper *et al.*[33] actually reported lower concordance levels for the 50% cut-off than for the 1% cut-off for assay 22C3 (overall percentage agreement (OPA) 81.9% and $\kappa = 0.58$ versus OPA 84.2% and $\kappa = 0.69$, respectively). However, this study reported prevalence bias to have influenced the κ magnitude for the 50% cut-off. These results therefore have to be interpreted with caution.

INTERLABORATORY CONCORDANCE

Interlaboratory concordance of TC staining was assessed by five of the 27 included studies (see Table 3; Table S7). Two of these[34,49] assessed only one antibody (22C3 and SP142, respectively). Both studies found high interlaboratory agreement. Adam *et al.*,[31] who assessed interlaboratory concordance for 22C3, 28-8 and SP263, found very high agreement between participating centres for each of these assays. Marchetti *et al.*[43] found similar results for the assays 22C3 and SP263. Scheel *et al.*[46] assessed interlaboratory concordance for standardised assays 22C3, 28-8, SP263 and SP142 and for 22C3, 28-8, SP263 and E1L3N used in LDTs, performed in 10 different sites. Concordance values ranged from Light's $\kappa = 0.63$–0.69 for the standardised assays when five cut-offs were used. κ was 0.49 for all the LDTs grouped together. When only a 1% and 50% cut-off were used, concordance values improved to $\kappa = 0.73$–0.89 for the standardised assays and $\kappa = 0.5$ for the LDTs.

Table 3. Summary of results from studies assessing interlaboratory concordance of TC scoring

Type of test	Interlaboratory concordance
Standardised assays	• 22C3: substantial to near-perfect concordance[31,34,43,46] • 28-8: substantial to near-perfect concordance[31,46] • SP263: substantial to near-perfect concordance[31,43,46] • SP142: high intersite percentage agreement[49]
LDTs	Only moderate concordance levels compared to standardised assays[46]

LDT, Laboratory-developed test; TC, Tumour cell.

CONCORDANCE OF IMMUNE CELL STAINING AND SCORING

A short analysis of concordance of IC staining and scoring can be found as Data S1 and Table S8.

Discussion

Ever since the approval of PD-1 and PD-L1 inhibitors as treatment options for patients with advanced NSCLC, various studies have been published assessing the comparability of different PD-L1 IHC assays. In this systematic review, interassay, interobserver and interlaboratory concordance of these PD-L1 IHC assays and LDTs were investigated by reviewing all currently available literature.

Overall, interassay agreement of TC staining is high between standardised assays 22C3, 28-8 and SP263, while assay SP142 frequently shows lower staining of TCs. Agreement between LDTs and their reference assay may also be high, depending on the protocol that is used, with some studies even showing greater agreement between LDTs and their reference assays than between different standardised assays.[28,44] These data seem to suggest that the assays 22C3, 28-8 and

SP263 and properly validated LDTs could be used interchangeably on histological specimens of NSCLC patients. However, some studies have shown lower concordance levels with the use of clinically relevant cut-offs.[22,28,44] The 1% cut-off especially may lead to higher disagreement compared to the 50% cut-off,[35,38,43] although this could perhaps be attributed to lower interrater agreement levels at this cut-off.[43] Based on the lower concordance levels found when using various cut-offs, it would be too premature to draw the conclusion that assays and LDTs can be used interchangeably without any consequences. Notably, in a recent meta-analysis of diagnostic accuracy of PD-L1 IHC assays, Torlakovic *et al.*[67] demonstrated that none of the standardised PD-L1 assays could be deemed as interchangeable, when interchangeability is defined as achieving ≥90% sensitivity and specificity for both the 1 and 50% cut-offs. Because discordance may exist between assays at clinically relevant cut-offs, simply interchanging one assay with another may potentially lead to patients being wrongfully denied valuable treatment options in clinical practice.

Assessment of interobserver concordance of TC scoring showed that agreement between pathologists is moderate to high for all assays and LDTs. Markedly, agreement is often found to be lowest for the 1% cut-off compared to other cut-offs. This is problematic, especially now that the European Medicines Agency (EMA) has only approved durvalumab as consolidation treatment in stage III NSCLC patients whose tumours show PD-L1 expression of ≥1%.[12] One could question if the use of this cut-off provides results that are reliable enough to aid clinicians in making treatment decisions. Agreement is likely to be higher between more experienced pathologists,[43] yet still leaves room for improvement. One study assessing training of already experienced pathologists showed no or only little improvement of interobserver agreement.[33] This study, however, employed a 1-h training session consisting of a presentation only. Alternative training initiatives, preferably including a more practical element during which trainees have to perform PD-L1 scoring on multiple specimens, might prove to be more effective. A recent study assessing interpathologist concordance of PD-L1 scoring using real-world data showed that training for PD-L1 scoring and experience in routine pathology practice correlated with higher concordance.[68] The effect of training on interobserver concordance should thus be studied more extensively. Other solutions also deserve more attention, especially the use of digital image analysis for PD-L1 scoring, as this has been shown to reduce interobserver variability.[69]

Finally, we assessed concordance of PD-L1 IHC assays and LDTs between laboratories. Only a limited number of studies assessed this type of concordance, especially compared to the large number of studies assessing interassay and/or interobserver concordance. Most of the studies assessing interlaboratory concordance found high agreement for all standardised assays, while one study found lower agreement for LDTs.[46] However, not all these studies used the right study protocol and the right outcome measure to properly assess interlaboratory concordance. Two studies[34,49] used percentage of agreement as outcome measure, which does not account for random agreement and may thus overestimate true agreement.[65,66] Two other studies[43,46] used study designs that did not allow for separate analysis of interobserver and interlaboratory concordance. Moreover, none of the study designs allowed for assessment of the influence of pre-analytical variables on PD-L1 immunostaining, while in clinical practice pre-analytical processing of samples may actually differ considerably between laboratories and may influence IHC staining results.[70-72] Therefore, studies assessing interlaboratory variation in PD-L1 expression are needed, using real-world data and thereby taking into account these possible differences in pre-analytical variables in clinical practice.

This systematic review has some limitations. Most importantly, there is significant heterogeneity between the studies included, especially in the choice of antibodies tested and the statistical methods used to analyse concordance. This prohibits pooling of data and complicates proper comparison of results between studies. Most studies, however, used similar samples for PD-L1 testing, i.e. formalin-fixed paraffin-embedded material from tumour resections or biopsies from NSCLC patients. This supports comparability between studies. Conversely, this also provides a disadvantage: it only allows for comparison of PD-L1 IHC assays and LDTs in histology, while in clinical practice PD-L1 immunostaining is frequently performed on cytological specimens. Comparison of PD-L1 IHC assays and LDTs in cytological NSCLC specimens falls beyond the scope of this review, but would be worth evaluation in a separate study. Finally, many of the included studies were not of high methodological quality, with only five studies being judged as having low risk of bias. Excluding the studies with the highest level of risk of bias, however, has improved the overall quality of this review.

To conclude, this systematic review has shown that interassay concordance of TC staining is generally high between the standardised assays 22C3, 28-8 and SP263 and properly developed and validated

LDTs. Nevertheless, the use of clinically relevant cut-offs may lead to lower levels of interassay concordance, indicating that these assays and LDTs cannot simply be interchanged. Interobserver agreement, moreover, is generally high for all assays and LDTs, but decreases with use of the 1% cut-off. Lastly, interlaboratory concordance seems to be high for standardised assays and moderate for LDTs, but has not been studied sufficiently to draw definitive conclusions. Studies using real-world clinical pathology data are necessary to assess whether use of different PD-L1 IHC assays and LDTs, scoring by different pathologists and use of different pre-analytical variables actually lead to differences in PD-L1 positivity between laboratories in clinical practice.

Acknowledgements

No grants or other forms of support were received for this study.

Conflicts of interest

B. K. and S. W. received research grants from AstraZeneca, MSD and Roche Diagnostics. S. W. also received research grants from BMS, NextCure and Pfizer. M. H. received research grants from AstraZeneca, BMS, MSD, Pfizer, Roche and Roche Diagnostics. S. B. declares no conflicts of interest. None of the grant suppliers were involved in the study design, collection, analysis and interpretation of data, writing of the manuscript or the decision to submit the manuscript for publication.

References

1. Li Y, Li F, Jiang F *et al.* A mini-review for cancer immunotherapy: molecular understanding of PD-1/PD-L1 pathway & translational blockade of immune checkpoints. *Int. J. Mol. Sci.* 2016; **17**; pii: E1151.
2. US Food and Drug Administration. YERVOY (ipilimumab) [highlights of prescribing information] [internet]. 2018. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/125377s094lbl.pdf (Accessed 30 October 2019).
3. European Mdicines Agency (EMA). Yervoy (ipilimumab) [internet]. 2019. Available at: https://www.ema.europa.eu/en/medicines/human/EPAR/yervoy (Accessed 30 October 2019).
4. Brahmer J, Reckamp KL, Baas P *et al.* Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N. Engl. J. Med.* 2015; **373**; 123–135.
5. 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**; 1627–1639.
6. Rittmeyer A, Barlesi F, Waterkamp D *et al.* Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet* 2017; **389**; 255–265.
7. Antonia SJ, Villegas A, Daniel D *et al.* Durvalumab after chemoradiotherapy in Stage III non-small-cell lung cancer. *N. Engl. J. Med.* 2017; **377**; 1919–1929.
8. Reck M, Rodriguez-Abreu D, Robinson AG *et al.* Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N. Engl. J. Med.* 2016; **375**; 1823–1833.
9. Herbst RS, Baas P, Kim DW *et al.* Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* 2016; **387**; 1540–1550.
10. Planchard D, Popat S, Kerr K *et al.* Metastatic non-small-cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 2018; **29**(Suppl 4): iv192–iv237.
11. US Food and Drug Administration. KEYTRUDA (pembrolizumab) [prescribing information] [internet]. 2018. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/125514s034lbl.pdf (Accessed 13 February 2019).
12. European Medicines Agency (EMA). Imfinzi (durvalumab) [internet]. 2018. Available at: https://www.ema.europa.eu/en/medicines/human/EPAR/imfinzi (Accessed 9 April 2019).
13. US Food and Drug Administration. Dako PD-L1 IHC 22C3 pharmDx [internet]. 2015. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf15/P150013c.pdf (Accessed 13 February 2019).
14. Dako Agilent Pathology Solutions. PD-L1 IHC 22C3 pharmDx is CE-IVD-marked for in vitro diagnostic use [internet]. Available at: https://www.agilent.com/cs/library/usermanuals/public/29171_22C3-ihc-pharmdx-interpretation-manual-eu.pdf (Accessed 30 October 2019).
15. US Food and Drug Administration. Dako PD-L1 IHC 28–8 pharmDx [internet]. 2015. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf15/P150025c.pdf (accessed 13 February 2019).
16. Dako Agilent Pathology Solutions. PD-L1 IHC 28–8 pharmDx product information. Available at: https://www.agilent.com/cs/library/brochures/29125_pd-l1-ihc-28-8-pharmdx-brochure_row.pdf (Accessed 30 October 2019).
17. Ventana Medical Systems Inc. VENTANA PD-L1 (SP142) assay [internet]. 2019. Available at: https://diagnostics.roche.com/global/en/products/tests/ventana-pd-l1-_sp142-assay2.html (Accessed 13 February 2019).
18. US Food and Drug Administration. VENTANA PD-L1 (SP142) assay. 2016. Available at: https://www.accessdata.fda.gov/cdrh_docs/pdf16/P160002c.pdf (Accessed 30 October 2019).
19. Buttner R, Gosney JR, Skov BG *et al.* Programmed death-ligand 1 immunohistochemistry testing: a review of analytical assays and clinical implementation in non-small-cell lung cancer. *J. Clin. Oncol.* 2017; **35**; 3867–3876.
20. Ventana Medical Systems, Inc. VENTANA PD-L1 (SP263) assay [internet]. 2019. Available at: https://diagnostics.roche.com/global/en/products/tests/ventana-pd-l1-_sp263-assay2.html (Accessed 13 February 2019).
21. Kerr KM, Hirsch FR. Programmed death ligand-1 immunohistochemistry: friend or foe? *Arch. Pathol. Lab. Med.* 2016; **140**; 326–331.
22. 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.

23. Ratcliffe MJ, Sharpe A, Midha A *et al*. Agreement between programmed cell death ligand-1 diagnostic assays across multiple protein expression cutoffs in non-small-cell lung cancer. *Clin. Cancer Res.* 2017; **23**; 3585–3591.
24. Rimm DL, Han G, Taube JM *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**; 1051–1058.
25. Scheel AH, Dietel M, Heukamp LC *et al*. Harmonized PD-L1 immunohistochemistry for pulmonary squamous-cell and adenocarcinomas. *Mod. Pathol.* 2016; **29**; 1165–1172.
26. Whiting PF, Rutjes AW, Westwood ME *et al*. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* 2011; **155**; 529–536.
27. Hayden JA, van der Windt DA, Cartwright JL, Cote P, Bombardier C. Assessing bias in studies of prognostic factors. *Ann. Intern. Med.* 2013; **158**; 280–286.
28. Hendry S, Byrne DJ, Wright GM *et al*. Comparison of four PD-L1 immunohistochemical assays in lung cancer. *J. Thorac. Oncol.* 2018; **13**; 367–376.
29. Russell-Goldman E, Kravets S, Dahlberg SE, Sholl LM, Vivero M. Cytologic–histologic correlation of programmed death-ligand 1 immunohistochemistry in lung carcinomas. *Cancer Cytopathol.* 2018; **126**; 253–263.
30. Skov BG, Skov T. Paired comparison of PD-L1 expression on cytologic and histologic specimens from malignancies in the lung assessed with PD-L1 IHC 28–8pharmDx and PD-L1 IHC 22C3pharmDx. *Appl. Immunohistochem. Mol. Morphol.* 2017; **25**; 453–459.
31. Adam J, Le Stang N, Rouquette I *et al*. Multicenter harmonization study for PD-L1 IHC testing in non-small-cell lung cancer. *Ann. Oncol.* 2018; **29**; 953–958.
32. Brunnström H, Johansson A, Westbom-Fremer S *et al*. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher than assay variability. *Mod. Pathol.* 2017; **30**; 1411–1421.
33. Cooper WA, Russell PA, Cherian M *et al*. Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non-small-cell lung cancer. *Clin. Cancer Res.* 2017; **23**; 4569–4577.
34. Roach C, Zhang N, Corigliano E *et al*. Development of a companion diagnostic PD-L1 immunohistochemistry assay for pembrolizumab therapy in non-small-cell lung cancer. *Appl. Immunohistochem. Mol. Morphol.* 2016; **24**; 392–397.
35. Chan AWH, Tong JHM, Kwan JSH *et al*. Assessment of programmed cell death ligand-1 expression by 4 diagnostic assays and its clinicopathological correlation in a large cohort of surgical resected non-small-cell lung carcinoma. *Mod. Pathol.* 2018; **31**; 1381–1390.
36. Cogswell J, Inzunza HD, Wu Q *et al*. An analytical comparison of Dako 28–8 PharmDx assay and an E1L3N laboratory-developed test in the immunohistochemical detection of programmed death-ligand 1. *Mol. Diagn. Ther.* 2017; **21**; 85–93.
37. Conde E, Caminoa A, Dominguez C *et al*. Aligning digital CD8 (+) scoring and targeted next-generation sequencing with programmed death ligand 1 expression: a pragmatic approach in early-stage squamous cell lung carcinoma. *Histopathology* 2018; **72**; 270–284.
38. Fujimoto D, Sato Y, Uehara K *et al*. Predictive performance of four programmed cell death ligand 1 assay systems on nivolumab response in previously treated patients with non-small-cell lung cancer. *J. Thorac. Oncol.* 2018; **13**; 377–386.
39. Ilie M, Falk AT, Butori C *et al*. PD-L1 expression in basaloid squamous cell lung carcinoma: relationship to PD-1(+) and CD8(+) tumor-infiltrating T cells and outcome. *Mod. Pathol.* 2016; **29**; 1552–1564.
40. Ilie M, Khambata-Ford S, Copie-Bergman C *et al*. Use of the 22C3 anti-PD-L1 antibody to determine PD-L1 expression in multiple automated immunohistochemistry platforms. *PLoS One* 2017; **12**; e0183023.
41. Ilie M, Juco J, Huang L, Hofman V, Khambata-Ford S, Hofman P. Use of the 22C3 anti-programmed death ligand 1 antibody to determine programmed death ligand 1 expression in cytology samples obtained from non-small-cell lung cancer patients. *Cancer Cytopathol.* 2018; **126**; 264–274.
42. Keller MD, Neppl C, Irmak Y *et al*. Adverse prognostic value of PD-L1 expression in primary resected pulmonary squamous cell carcinomas and paired mediastinal lymph node metastases. *Mod. Pathol.* 2018; **31**; 101–110.
43. Marchetti A, Barberis M, Franco R *et al*. Multicenter comparison of 22C3 PharmDx (Agilent) and SP263 (Ventana) assays to test PD-L1 expression for NSCLC patients to be treated with immune checkpoint inhibitors. *J. Thorac. Oncol.* 2017; **12**; 1654–1663.
44. Munari E, Rossi G, Zamboni G *et al*. PD-L1 assays 22C3 and SP263 are not interchangeable in non-small-cell lung cancer when considering clinically relevant cutoffs: an interclone evaluation by differently trained pathologists. *Am. J. Surg. Pathol.* 2018; **42**; 1384–1389.
45. Rehman JA, Han G, Carvajal-Hausdorf DE *et al*. Quantitative and pathologist-read comparison of the heterogeneity of programmed death-ligand 1 (PD-L1) expression in non-small-cell lung cancer. *Mod. Pathol.* 2017; **30**; 340–349.
46. Scheel AH, Baenfer G, Baretton G *et al*. Interlaboratory concordance of PD-L1 immunohistochemistry for non-small-cell lung cancer. *Histopathology* 2018; **72**; 449–459.
47. Soo RA, Lim JSY, Asuncion BR *et al*. Determinants of variability of five programmed death ligand-1 immunohistochemistry assays in non-small-cell lung cancer samples. *Oncotarget* 2018; **9**; 6841–6851.
48. Tsao MS, Kerr KM, Kockx M *et al*. PD-L1 immunohistochemistry comparability study in real-life clinical samples: results of Blueprint Phase 2 project. *J. Thorac. Oncol.* 2018; **13**; 1302–1311.
49. Vennapusa B, Baker B, Kowanetz M *et al*. Development of a PD-L1 complementary diagnostic immunohistochemistry assay (SP142) for atezolizumab. *Appl. Immunohistochem. Mol. Morphol.* 2019; **27**; 92–100.
50. Xu H, Lin G, Huang C *et al*. Assessment of concordance between 22C3 and SP142 immunohistochemistry assays regarding PD-L1 expression in non-small-cell lung cancer. *Sci Rep.* 2017; **7**; 16956.
51. Erber R, Stohr R, Herlein S *et al*. Comparison of PD-L1 mRNA expression measured with the CheckPoint Typer(R) assay with PD-L1 protein expression assessed with immunohistochemistry in non-small-cell lung cancer. *Anticancer Res.* 2017; **37**; 6771–6778.
52. Kim H, Kwon HJ, Park SY, Park E, Chung JH. PD-L1 immunohistochemical assays for assessment of therapeutic strategies involving immune checkpoint inhibitors in non-small-cell lung cancer: a comparative study. *Oncotarget* 2017; **8**; 98524–98532.
53. Krawczyk P, Jarosz B, Kucharczyk T *et al*. Immunohistochemical assays incorporating SP142 and 22C3 monoclonal antibodies for detection of PD-L1 expression in NSCLC patients

with known status of EGFR and ALK genes. *Oncotarget* 2017; **8**; 64283–64293.
54. McLaughlin J, Han G, Schalper KA *et al*. Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer. *JAMA Oncol.* 2016; **2**; 46–54.
55. Neuman T, London M, Kania-Almog J *et al*. A Harmonization study for the use of 22C3 PD-L1 immunohistochemical staining on Ventana's platform. *J. Thorac. Oncol.* 2016; **11**; 1863–1868.
56. Pang C, Yin L, Zhou X *et al*. Assessment of programmed cell death ligand-1 expression with multiple immunohistochemistry antibody clones in non-small-cell lung cancer. *J. Thorac. Dis.* 2018; **10**; 816–824.
57. Parra ER, Villalobos P, Mino B, Rodriguez-Canales J. Comparison of different antibody clones for immunohistochemistry detection of programmed cell death ligand 1 (PD-L1) on non-small-cell lung carcinoma. *Appl. Immunohistochem. Mol. Morphol.* 2018; **26**; 83–93.
58. Paulsen EE, Kilvaer TK, Khanehkenari MR *et al*. Assessing PDL-1 and PD-1 in non-small-cell lung cancer: a novel immunoscore approach. *Clin. Lung Cancer* 2017; **18**; 220–233.e8.
59. 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**; 541–549.
60. 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.
61. Roge R, Vyberg M, Nielsen S. Accurate PD-L1 protocols for non-small-cell lung cancer can be developed for automated staining platforms with clone 22C3. *Appl. Immunohistochem. Mol. Morphol.* 2017; **25**; 381–385.
62. Sheffield BS, Fulton R, Kalloger SE *et al*. Investigation of PD-L1 biomarker testing methods for PD-1 axis inhibition in non-squamous non-small-cell lung cancer. *J. Histochem. Cytochem.* 2016; **64**; 587–600.
63. Smith J, Robida MD, Acosta K *et al*. Quantitative and qualitative characterization of two PD-L1 clones: SP263 and E1L3N. *Diagn. Pathol.* 2016; **11**; 44.
64. Tseng JS, Yang TY, Wu CY *et al*. Characteristics and predictive value of PD-L1 status in real-world non-small cell lung cancer patients. *J. Immunother.* 2018; **41**; 292–299.
65. Mandrekar JN. Measures of interrater agreement. *J. Thorac. Oncol.* 2011; **6**; 6–7.
66. McHugh ML. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)* 2012; **22**; 276–282.
67. Torlakovic E, Lim HJ, Adam J *et al*. 'Interchangeability' of PD-L1 immunohistochemistry assays: a meta-analysis of diagnostic accuracy. *Mod. Pathol.* 2019. https://doi.org/10.1038/s41379-019-0327-4.
68. Adam J, Hofman V, Mansuet-Lupo A *et al*. P2.09-17 real-world concordance across pathologists for PD-L1 scoring in non-small-cell lung cancer: results from a large nationwide initiative. *J. Thorac. Oncol.* 2019; **14**; S775.
69. Koelzer VH, Gisler A, Hanhart JC *et al*. Digital image analysis improves precision of PD-L1 scoring in cutaneous melanoma. *Histopathology* 2018; **73**; 397–406.
70. Khoury T, Sait S, Hwang H *et al*. Delay to formalin fixation effect on breast biomarkers. *Mod. Pathol.* 2009; **22**; 1457–1467.
71. Babic A, Loftin IR, Stanislaw S *et al*. The impact of pre-analytical processing on staining quality for H&E, dual hapten, dual color in situ hybridization and fluorescent *in situ* hybridization assays. *Methods* 2010; **52**; 287–300.
72. Wolff AC, Hammond ME, Hicks DG *et al*. Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin. Oncol.* 2013; **31**; 3997–4013.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Search syntax in PubMed, Embase and Cochrane Library.

Table S2. Inclusion and exclusion criteria.

Table S3. Quality assessment of included studies.

Table S4. Study characteristics of all studies included for data extraction and analysis.

Table S5. Results from studies assessing inter-assay concordance of TC staining.

Table S6. Results from studies assessing inter-observer concordance of TC scoring.

Table S7. Results from studies assessing inter-laboratory concordance of TC staining.

Table S8. Results from studies assessing inter-assay and/or inter-observer concordance of IC staining/scoring.

Data S1. Supplementary results: concordance of IC staining and scoring.