Research Article

Multi-Field-of-View Deep Learning Model Predicts Nonsmall Cell Lung Cancer Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images

Lingdao Sha[1], Boleslaw L. Osinski[1], Irvin Y. Ho[1], Timothy L. Tan[1,2], Caleb Willis[1], Hannah Weiss[1,3], Nike Beaubier[1], Brett M. Mahon[1], Tim J. Taxter[1], Stephen S. F. Yip[1]

[1]Tempus Labs, Inc, Chicago, IL USA, [2]Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA, [3]Department of Neurological Surgery, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Received: 26 April 2019 Accepted: 06 June 2019 Published: 23 July 2019

Abstract

Background: Tumor programmed death-ligand 1 (PD-L1) status is useful in determining which patients may benefit from programmed death-1 (PD-1)/PD-L1 inhibitors. However, little is known about the association between PD-L1 status and tumor histopathological patterns. Using deep learning, we predicted PD-L1 status from hematoxylin and eosin (H and E) whole-slide images (WSIs) of nonsmall cell lung cancer (NSCLC) tumor samples. **Materials and Methods:** One hundred and thirty NSCLC patients were randomly assigned to training (n = 48) or test (n = 82) cohorts. A pair of H and E and PD-L1-immunostained WSIs was obtained for each patient. A pathologist annotated PD-L1 positive and negative tumor regions on the training samples using immunostained WSIs for reference. From the H and E WSIs, over 145,000 training tiles were generated and used to train a multi-field-of-view deep learning model with a residual neural network backbone. **Results:** The trained model accurately predicted tumor PD-L1 status on the held-out test cohort of H and E WSIs, which was balanced for PD-L1 status (area under the receiver operating characteristic curve [AUC] =0.80, $P << 0.01$). The model remained effective over a range of PD-L1 cutoff thresholds (AUC = 0.67–0.81, $P \leq 0.01$) and when different proportions of the labels were randomly shuffled to simulate interpathologist disagreement (AUC = 0.63–0.77, $P \leq 0.03$). **Conclusions:** A robust deep learning model was developed to predict tumor PD-L1 status from H and E WSIs in NSCLC. These results suggest that PD-L1 expression is correlated with the morphological features of the tumor microenvironment.

Keywords: Artificial intelligence, deep learning, digital pathology, lung cancer

Introduction

Nonsmall cell lung cancer (NSCLC) is the most common type of lung cancer, affecting over 1.5 million people worldwide.[1] The disease often responds poorly to standard of care chemoradiotherapy and has a high incidence of recurrence, resulting in low 5-year survival rates.[2-4] Advances in immunology showed that NSCLC frequently elevates the expression of programmed death-ligand 1 (PD-L1) to bind to programmed death-1 (PD-1) expressed on the surface of T-cells.[5,6] PD-1 and PD-L1 binding deactivates T-cell antitumor responses, enabling NSCLC to evade targeting by the immune system.[7] The discovery of the interplay between tumor progression and immune response has led to the development and regulatory approval of PD-1/PD-L1 checkpoint blockade immunotherapies such as nivolumab and pembrolizumab.[8-10] Anti-PD-1 and anti-PD-L1 antibodies restore antitumor immune response by disrupting the interaction between PD-1 and PD-L1.[11] Notably, PD-L1-positive NSCLC patients treated with these checkpoint inhibitors achieve durable tumor regression and improved survival.[12-16]

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_24_19

Address for correspondence: Dr. Stephen S. F. Yip, Tempus Labs, Inc., 600 West Chicago Ave. Ste 510, Chicago, IL 60608, USA.
E-mail: stephen.yip@tempus.com

How to cite this article: Sha L, Osinski BL, Ho IY, Tan TL, Willis C, Weiss H, *et al*. Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images. J Pathol Inform 2019;10:24.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2019/10/1/24/263339

As the role of immunotherapy in oncology expands, it is useful to accurately assess tumor PD-L1 status to identify patients who may benefit from PD-1/PD-L1 checkpoint blockade immunotherapy. Immunohistochemistry (IHC) staining of tumor tissues acquired from biopsy or surgical specimens is commonly employed to assess PD-L1 status.[17-19] However, IHC staining can be limited by insufficient tissue samples, and in some settings, a lack of resources.[20,21]

Hematoxylin and eosin (H and E) staining is fundamental to analyzing tissue morphological features for malignancy diagnosis, including NSCLC.[22,23] Furthermore, H and E slides may capture tissue visual characteristics that are associated with PD-L1 status. For example, Velcheti *et al*. and McLaughlin *et al*. both observed that PD-L1 positive NSCLC tended to have higher levels of tumor-infiltrating lymphocytes (TILs).[24,25] However, quantification of TILs using H and E slides is laborious and affected by interobserver variability.[25,26] Moreover, TILs may be inadequate to fully describe the complexity of the tumor microenvironment and its relationship with PD-L1 status.

Technological advances have enabled the digitization of histopathology H and E and IHC slides into high-resolution whole-slide images (WSIs), providing opportunities to develop computer vision tools for a wide range of clinical applications.[27-29] Recently, deep learning applications to pathology images have shown tremendous promise in predicting treatment outcomes,[30] disease subtypes,[31,32] lymph node status,[27,28] and genetic characteristics[30,33,34] in various malignancies. Deep learning is a subset of machine learning wherein models are built with a number of discrete neural node layers, imitating the structure of the human brain.[35] These models learn to recognize complex visual features from WSIs by iteratively updating the weighting of each neural node based on the training examples.[29] To our knowledge, deep learning has never been applied to predicting PD-L1 status from H and E imaging features in NSCLC patients.

Here, we hypothesized that morphological changes to the NSCLC microenvironment associated with elevated PD-L1 expression can be recognized by deep learning. Our results may aid the further development of H and E -based imaging biomarkers that complement clinical IHC testing for tumor PD-L1 status.

Materials and Methods

The tissue samples and patient-related health information used in the study that is described in the manuscript were deidentified. All digital images and clinical data related to the tumor tissues were also anonymized.

Programmed death-ligand 1 immunohistochemistry assay

A total of 130 deidentified, archival formalin-fixed, paraffin-embedded (FFPE) tumor tissues (1 tissue sample/patient) from NSCLC patients were used in this study. All the FFPE blocks were processed, reviewed, and stored in a College of American Pathologists (CAP) accredited and Clinical Laboratory Improvement Amendments (CLIA) certified laboratory (Tempus Labs Chicago, IL, USA).

Each FFPE block was cut into 4 μm-thick serial sections for H and E and IHC stains. H and E staining was performed on the Leica Autostainer XL staining platform. The sections were stained with H and E and anti-PD-L1 (clone 22c3, pharmDx Kit Dako) using an automated staining system (BOND-III: Leica Microsystems). IHC slides were stained with anti-PD-L1 22C3 monoclonal mouse primary antibody using the Bond Polymer Refine detection system on a Leica Microsystems BOND-III with positive and negative cell line run controls.

Assessment of programmed death-ligand 1 expression

The level of PD-L1 expression, referred to as tumor PD-L1 score or tumor proportion score, was defined as the number of partially or completely stained tumor cells at any intensity divided by the total number of tumor cells.[36] In accordance with the Food and Drug Administration (FDA) documents, a tumor tissue was considered to have a PD-L1 positive (PD-L1+) status if its expression level was >1%.[15,36] Tumor tissue with a PD-L1 expression level ≤1% was considered PD-L1 negative (PD-L1−). One of three CAP-CLIA certified laboratory pathologists (N.B., B.M., or T.J.T.) reviewed IHC-stained slides and scored the level of PD-L1 expression.

Dataset distribution

A total of 130 H and E and their corresponding PD-L1 IHC slides were scanned and digitized at a resolution of 0.25 μm/pixel (40× magnification) using a Philips Ultra Fast Scanner (Philips, Eindhoven, The Netherlands). The images were initially acquired in the iSyntax format and then converted to the tagged image format file (TIFF) format using Philips' proprietary algorithm. Eighty-two of these WSIs were randomly chosen as an independent test cohort. To ensure a balanced test cohort (i.e., 41 PD-L1+ and 41 PD-L1−), these slides were equally sampled from tumor PD-L1+ and PD-L1− cases [Table 1]. The remaining 48 were used as a training cohort to train our deep learning architecture [Figure 1a and b].

Training example generation

In the training cohort, PD-L1+ and PD-L1− tumor regions of IHC slides were manually annotated by a pathologist (T.L.T.). Annotations were made using the publicly-available digital pathology software QuPath.[37] Then, using the IHC annotations as a reference, matching areas on the corresponding H and E slides were annotated [Figure 1b]. In total, three classes (i.e., tumor PD-L1+, tumor PD-L1−, and other) were annotated on the H and E slides. Nontumor regions on the H and E slides, such as stroma, necrosis, and normal epithelium, were annotated as a single "other" class.

The annotated regions were tiled into overlapping tiles (466 × 466 pixels) with a stride of 32 pixels at 10× magnification (1 pixel = 1 μm). Only tiles whose center fell within the annotated regions were kept. The tile size was chosen to provide spatial context to the network, while the stride was chosen to increase sampling density and the number

Table 1: Patient characteristics in the test and training cohorts

	Test cohort			Training cohort		
	PD-L1+ (*n*=41), *n* (%)	PD-L1− (*n*=41), *n* (%)	Overall (*n*=82), *n* (%)	PD-L1+ (*n*=28), *n* (%)	PD-L1− (*n*=20), *n* (%)	Overall (*n*=48), *n* (%)
Age (year)						
Average	70	73	72	70	68	69
Range	38-93	57-86	38-93	50-87	36-84	36-87
Sex						
Male	26 (63)	17 (41)	43 (52)	12 (43)	13 (65)	25 (52)
Female	15 (37)	24 (59)	39 (48)	16 (57)	7 (35)	23 (48)
Smoking history						
Current/former smoker	31 (76)	30 (73)	61 (74)	21 (75)	13 (65)	34 (71)
Never smoker	4 (10)	7 (17)	11 (13)	3 (11)	3 (15)	6 (13)
N/A	6 (15)	4 (10)	10 (12)	4 (14)	4 (20)	8 (17)
Overall stages						
IA/IB	11 (27)	16 (39)	27 (33)	7 (25)	8 (40)	15 (31)
IIA/IIB	6 (15)	11 (27)	17 (21)	5 (18)	2 (10)	7 (15)
IIIA/IIIB	10 (24)	7 (17)	17 (21)	3 (11)	2 (10)	5 (10)
IV	8 (20)	6 (15)	14 (17)	11 (39)	8 (40)	19 (40)
N/A	6 (15)	1 (2)	7 (8)	2 (7)	0	2 (4)
Histology subtypes						
Adenocarcinoma	30 (73)	31 (76)	61 (74)	20 (71)	17 (85)	37 (77)
SCC	7 (17)	10 (24)	17 (21)	7 (25)	3 (15)	10 (21)
Adenosquamous	4 (10)	0	4 (5)	1 (4)	0	1 (2)

Tumor PD-L1 + and PD-L1−status were determined using immunohistochemistry staining. N/A=Information not available, SCC=Squamous cell carcinoma, PD-L1=Programmed death-ligand 1, PD-L1+=PD-L1 positive, PD-L1−=PD-L1 negative

of training examples. This procedure produced 107,854 PD-L1+ and 57,837 PD-L1− tiles.

Deep learning architecture

Our deep learning architecture is composed of three major components: (1) a fully convolutional residual neural network (ResNet) backbone that processes a large 466 × 466 field of view (FOV), (2) two branches that process 32 × 32 small FOVs, and (3) concatenation of small and large FOV features for multi-FOV classification [Figure 1].

We chose ResNet because it overcomes the accuracy degradation challenges traditionally suffered by "very deep" neural networks (i.e., neural networks with more than 16 convolutional layers).[38,39] ResNet consists of a stack of convolutional layers interleaved with "shortcut connections" which skip intermediate layers [Figure 1a]. These connections use earlier layers as a reference point to guide deeper layers to learn the residual between layer outputs rather than learning an identity mapping between layers. This innovation improves convergence speed and stability during training and allows deeper networks to perform better than their shallower counterparts.[38]

The backbone of our model consisted of an 18-layer version of ResNet (ResNet-18) with some modifications. The ResNet-18 backbone was converted into a fully convolutional network (FCN) by removing the global average pooling layer and eliminating zero padding in downsampled layers. This enables the output of a two-dimensional probability map rather than a one-dimensional probability vector [Figure 1c]. The tile size (466 × 466 pixels) is over twice the tile size of a standard ResNet, providing our model with a larger FOV that allows it to learn surrounding morphological features.

The model includes two additional branches with receptive fields restricted to a small FOV (32 × 32 pixels) in the center of the second convolutional feature map [Figure 1b]. One branch passes a copy of the small FOV through a convolutional filter, while the other branch is a standard shortcut connection with downsampling. The features produced by these additional branches are concatenated to the features from the main backbone just before the model outputs are converted into probabilities in the softmax layer. In this way, the model combines information from multiple FOVs, much like a pathologist relies on various zoom levels when diagnosing slides; our implementation ensures that the central region of each tile contributes more to classification than the tile edges, resulting in a more accurate classification map across the entire WSI.

Implementation details

Our model was implemented using PyTorch and trained on an NVIDIA Tesla V100 GPU by stochastic gradient descent with a batch size of 100. The ResNet-18 backbone was initialized with pretrained ImageNet weights.[40] Image augmentations, including random crop, random rotation, random flip, and color jitter, were performed batchwise during training. Batch normalization,[41]

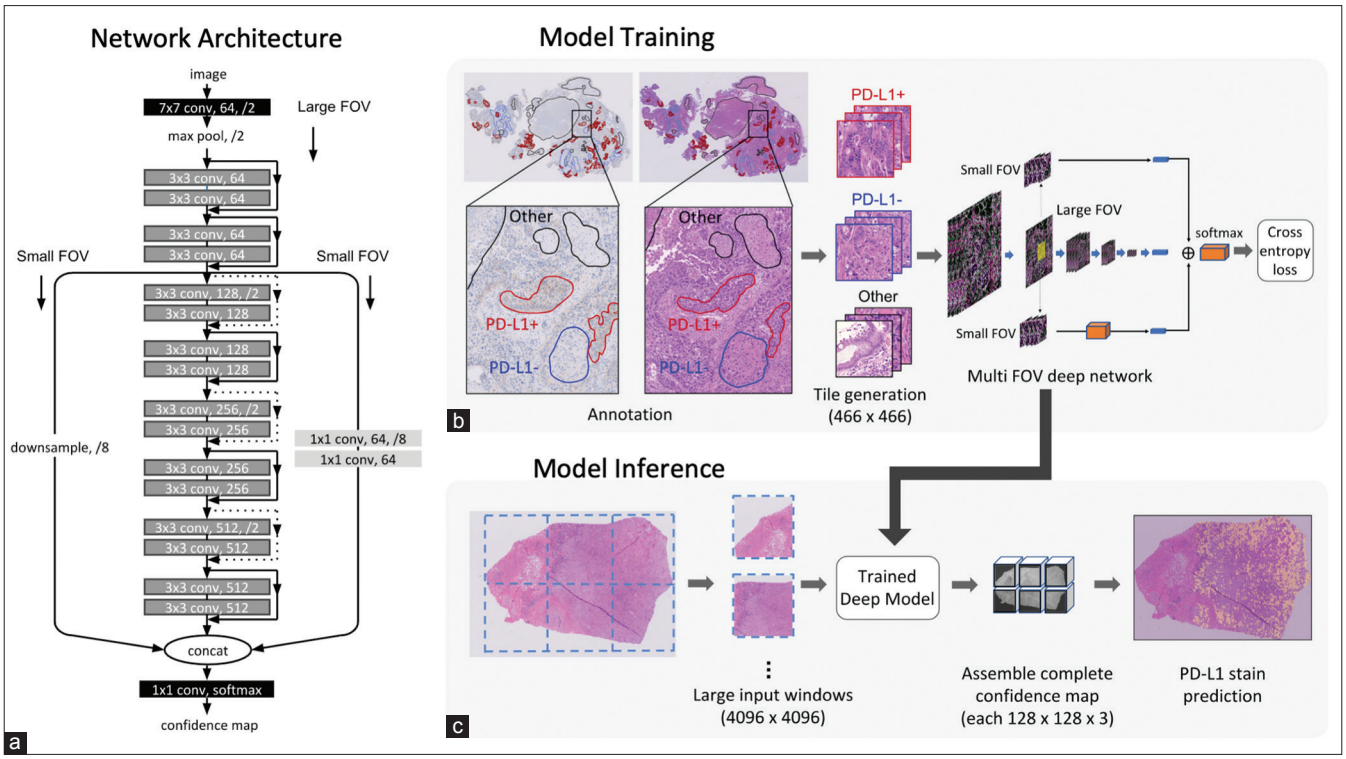

Figure 1: (a) Network architecture: Our deep learning framework consists of a fully convolutional ResNet-18 that processes a large field of view, along with two additional branches that process small field of views. The ResNet-18 backbone contains multiple shortcut connections. The dotted lines indicate shortcut connections where feature maps are also downsampled by 2. The small field-of-view branches emerge after the second convolutional block. The feature maps of the small field-of-view branches are downsampled by 8 to match the dimensions of the ResNet-18 feature map. These feature maps are concatenated before passing through a softmax output to produce a programmed death-ligand 1 staining probability map. (b) Model training: matching areas on Immunohistochemistry and H and E slides were annotated. The annotated regions of the H and E image were tiled into overlapping tiles (466 × 466 pixels) with a stride of 32 pixels, producing our training data. The multi-field-of-view ResNet-18 model was then trained using a cross-entropy loss function. The yellow square in the model schematic depicts the central region that is cropped for the small field of views. (c) Model inference: each image was divided into large nonoverlapping 4096 × 4096 input windows (blue dashed lines). Each large window was passed through the trained model. Because the model is fully convolutional, each tile within the large input window was processed in parallel, producing a 128 × 128 × 3 probability cube (the last dimension represents three classes). The resulting probability cubes were slotted into place and assembled to generate a probability map of the whole image. The class with the maximum probability was assigned to each tile

where each layer is independently normalized by subtracting the mean and dividing by the standard deviation (SD) of each training batch, was also implemented to accelerate training, improve network stability, and reduce overfitting. We used a cross-entropy loss function with the Adam optimizer and an initial learning rate of 0.001, which was decreased by 50% at epochs 3, 5, 7, and 9. Twenty percentage of training tiles were randomly reserved for a validation set to monitor validation accuracy during training. The model was trained with 10 epochs, at which point validation loss no longer decreased.

Model inference

Inference on unseen WSIs was performed in three stages: (1) division of the image into large input windows consisting of many overlapping tiles, (2) simultaneous classification of all tiles within each input window using the trained fully-convolutional deep learning model, and (3) Computation of model score by aggregating tile classifications.

Since the tiles are 466 × 466 pixels large with a stride of only 32 pixels, they overlap significantly. While this overlap ensures that the resulting WSI classification map is smooth, it also represents a substantial potential for computational redundancy, as the convolutional features computed for overlapping regions will be identical. A single WSI can be over 5GB, with $\geq 10^{10}$ pixels and hundreds of thousands of tiles. To reduce this redundant computation, we took advantage of the fully convolutional nature of our model [Figure 1c]. Instead of using 466 × 466 pixel tiles as the model input, we used much larger 4096 × 4096 pixels input windows. The FCN treats the large input window as an array of tiles and computes their features simultaneously. Consequently, many of the convolutional features computed for a single tile are reused for neighboring tiles, enabling efficient computation.

Processing all tiles in this fashion produces a 2D probability map of the three classes (i.e., tumor PD-L1+, tumor PD-L1−, and other) for each tile in the image. The class with the highest probability is assigned to each tile. The model score is then calculated as the ratio of the number of predicted tumor PD-L1+ tiles to the total number of predicted tumor tiles (both tumor PD-L1+ and tumor PD-L1−). The model

score ranged from 0% to 100%, where 0% implies no tumor tiles are predicted to be PD-L1+ and 100% implies all tumor tiles are predicted to be PD-L1+.

Statistical analysis with an independent test cohort

Two analyses were performed to evaluate prediction of PD-L1 status based on H and E images alone. Both analyses were performed on the 82 "unseen" test cases that were held out from the model during training. In the first analysis, we assessed whether the average model score of the predicted PD-L1+ group was significantly greater than the PD-L1− group using the Welch's *t*-test (considered significant if $P_{\text{t-test}} < 0.05$). Average model score was defined as the score averaged over all patients within the group. For example, in the PD-L1+ group, the average model score was equal to the sum of each patient's model score divided by 42 (i.e., the total number of patients in the PD-L1+ group).

Second, the area under the receiver operating characteristic curve (AUC) was employed to quantify the power of the deep learning model score in predicting tumor PD-L1 status. A permutation test was implemented to assess if AUC was significantly different from random chance. In the permutation test, all PD-L1 labels of the test cohort were randomized 3000 times. A new AUC (AUC_{new}) was computed each time the label was randomized. AUC was considered to be significantly different from random if $AUC_{new} \geq AUC$ <5% of the time (i.e., $P_{permutation} < 0.05$). We reported $P_{permutation} << 0.01$ if none of the $AUC_{new} \geq AUC$.

Both analyses were also performed independently on adenocarcinoma and squamous cell carcinoma (SCC) cases.

Robustness studies

The robustness of our model was examined by testing the impact of different PD-L1 positivity cutoffs and shuffled PD-L1 statuses.

Programmed death-ligand 1 cutoff variation

The effect of changing the PD-L1 positivity cutoff on model predictions was investigated. For example, if the PD-L1 expression cutoff was set to 15%, a tumor tissue was only considered to be PD-L1+ if over 15% of the tumor cells were stained. The default cutoff in the test cohort was chosen to be 1% as aforementioned. Here, the cutoff was varied from 5% to 50% in increments of 5% and the AUC was computed for each cutoff.

Programmed death-ligand 1 label shuffle

To simulate the effect of pathologist variability, 5%–30% of the PD-L1 status labels in the test cohort were randomly shuffled in increments of 5%. Each shuffle was repeated 3000 times. The AUC was calculated for each repetition, and the permutation test assessed whether the average AUC (over 3000 repetitions) was significant.

Results

This study assessed the ability of our deep learning model to predict PD-L1 status from H and E images in 82 independent test cases. Most patients (>70%) in the test cohort had adenocarcinoma and were former/current smokers [Table 1]. Approximately half of the patients were female. Of all patients, 54% (44/82) and 38% (31/82) were overall Stage I/II and Stage III/IV, respectively. However, stage information was unavailable for 8% (7/82) patients. The test cohort was perfectly balanced, consisting of 50% (41/82) PD-L1+ and 50% (41/82) PD-L1− cases to prevent bias in our model evaluation.

The test WSIs ranged in size from 500MB to 5GB and measured 10,000–200,000 pixels in each dimension. The average computation time to generate a PD-L1 probability map was 40 s (7.9–66 s) on an AWS EC2 p3.2× large instance (NVIDIA V100, Intel Xeon E5-2686).

For PD-L1+ slides, the location of tissue regions with high predicted PD-L1+ probability typically corresponded to the location of observed PD-L1 IHC expression [Figure 2a-c]. For PD-L1− slides, most of the slide area was usually predicted to be tumor PD-L1− or other, with only a few sparse areas predicted to be tumor PD-L1+ [Figure 2d-f]. For all NSCLC test cases, the average deep learning model score for PD-L1+ WSIs (26% ±24%) was significantly greater than that for PD-L1− WSIs (6.5%±9.7%) with $P_{t\text{-test}} = 4.01 \times 10^{-7}$. As observed in Figure 3, the model score significantly separated PD-L1+ samples from PD-L1− samples. The deep learning model significantly predicted PD-L1 status in all test cases with an AUC of 0.80 ($P_{permutation} << 0.01$).

The analysis was also performed independently for the two NSCLC subtypes, lung adenocarcinoma and lung SCC. The model scores were greater for PD-L1+ than PD-L1− cases in both subtypes [Figure 3b-c]. The model score was observed to significantly discriminate PD-L1+ adenocarcinomas from PD-L1− adenocarcinomas (average model score = 28% ±25% vs. 5.3% ±7.5%) with $P_{t\text{-test}} = 1.6 \times 10^{-6}$ and AUC = 0.83 ($P_{permutation} << 0.01$). On the other hand, PD-L1+ SCC and PD-L1− SCC were not significantly discriminated by the model score in SCC ($P_{t\text{-test}} = 0.32$, AUC = 0.64, $P_{permutation} = 0.18$). The average model score was 23% ±27% and 10% ±14% for PD-L1+ SCC and PD-L1− SCC, respectively.

Robustness studies

Programmed death-ligand 1 cutoff variation

The impact of the PD-L1 positivity cutoff on the robustness of our model's predictive power was investigated. The AUCs varied moderately between 0.81 and 0.67 ($P_{permutation} \leq 0.01$) as the cutoff threshold used to determine positive PD-L1 status increased from 5% to 50% [Figure 4a]. Notably, the model score significantly separated PD-L1+ samples from PD-L1− samples for all cutoffs with $P_{t\text{-test}} \leq 0.03$.

Programmed death-ligand 1 label shuffle

The effect of interpathologist variability on the model was simulated by shuffling PD-L1 status label. The average

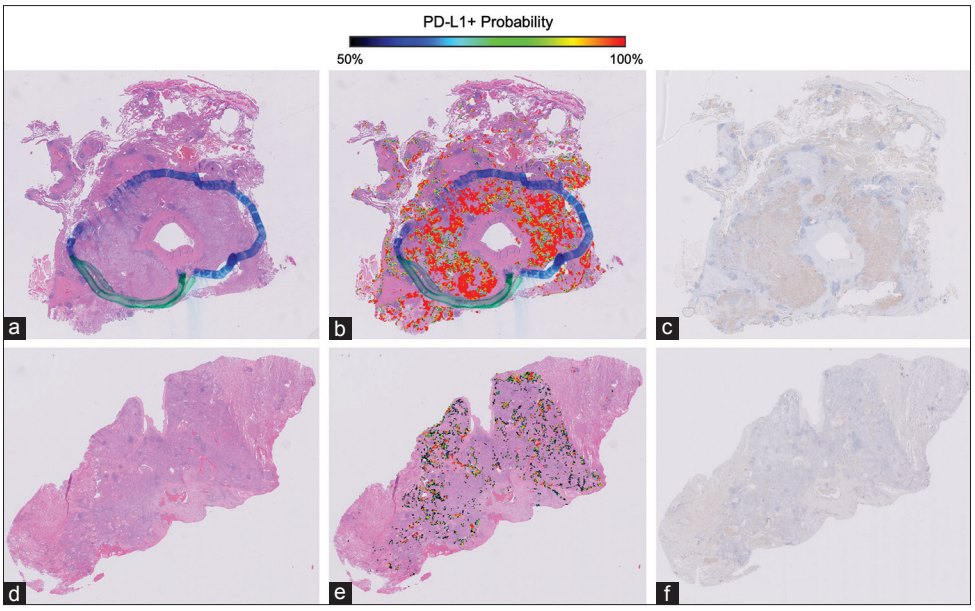

Figure 2: Top row: representative positive case (a) H and E whole-slide image, (b) probability map overlaid on H and E, and (c) programmed death-ligand 1 immunohistochemistry stain. Bottom row: representative negative case (d) H and E whole-slide image, (e) probability map overlaid on H and E, and (f) programmed death-ligand 1 immunohistochemistry stain. The color bar indicates the predicted probability of the tumor programmed death-ligand 1 + class. The outline marked in A and B is a laboratory remnant and unrelated to the model

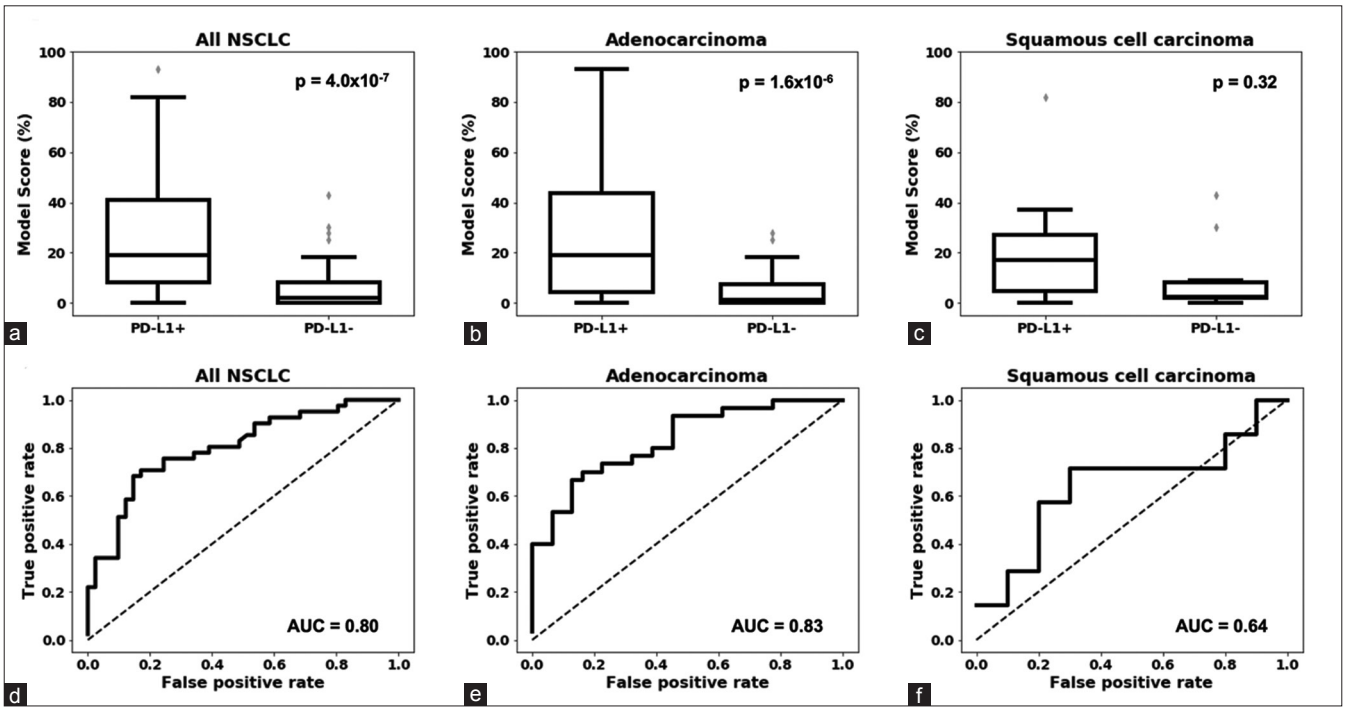

Figure 3: Test cohort results. Top row: Box plots depicting how tumor programmed death-ligand 1 statuses are separated by deep learning model score in (a) all nonsmall cell lung cancer, (b) lung adenocarcinoma, (c) lung squamous cell carcinoma. Bottom row: Receiver operating characteristic curve for (d) all nonsmall cell lung cancer, (e) adenocarcinoma, and (f) squamous cell carcinoma. The horizontal line indicates median

AUC decreased steadily from 0.77 ± 0.02 to 0.63 ± 0.05 as 5% (4/82) to 30% (25/82) of the labels were randomly shuffled [Figure 4b]. The overall SDs of the average AUC ranged from 0.02 to 0.08. All of the averaged AUCs were significant ($P_{permutation} \leq 0.03$).

Discussion

In this study, we developed and trained a deep learning model to recognize H and E imaging patterns for tumor PD-L1 status prediction in NSCLC. During training, our model was presented with over 145,000 examples to learn H and E image

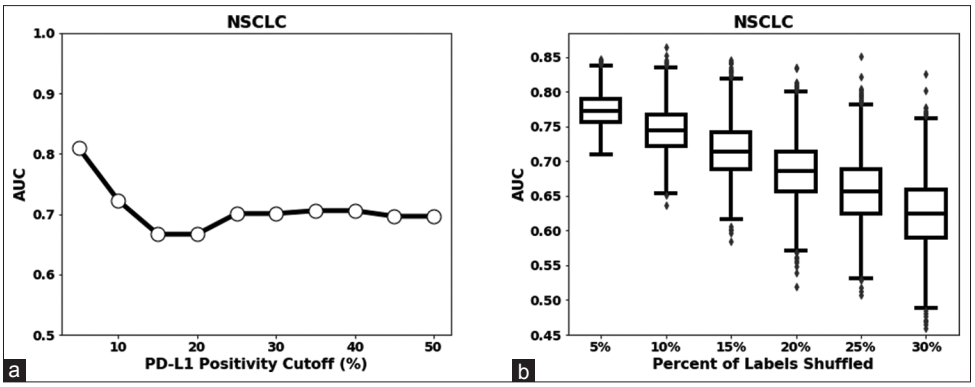

Figure 4: (a) Area under the receiver operating characteristic curve as a function of the programmed death-ligand 1 positivity cutoffs (b) area under the receiver operating characteristic curve as a function of the percentage of shuffled programmed death-ligand 1 labels

characteristics of PD-L1+ and PD-L1− tissue. The trained model reliably predicted PD-L1 status in the "unseen" test cohort (AUC = 0.80). Using ResNet18 as the backbone, our trained multi-FOV deep learning model reliably predicted PD-L1 status in the "unseen" test cohort (AUC = 0.80). Other architectures, such as ResNet 50, have been shown to outperform ResNet18 in image classification tasks.[38] Replacing our deep learning model backbone with other existing architectures (e.g., ResNet50, Inception v3) may further improve the ability of deep learning to predict PD-L1 status in NSCLC. The effect of backbone architecture on performance needs to be further investigated.

A convolutional neural network like our model learns a series of features to classify images. These features can be subtle and imperceptible to the naked eye. This method is distinct from previous studies that sought to recognize associations between tumor PD-L1 status and specific pathologist-defined features. For example, an increased density of TILs has been associated with PD-L1+ status in multiple malignancies.[25,42-44] However, manual quantification of TILs on WSIs is subjective and time-consuming. Furthermore, the microenvironment driven by the interaction between a tumor and the immune system is highly complex, and therefore, high levels of TILs and PD-L1 expression may not always co-occur.[5] Thus, a deep learning model that directly predicts PD-L1 status from imaging may represent a more holistic approach. Future work should investigate the relationship between TIL density and the deep learning features that are important for PD-L1 prediction.

Our deep learning model predicted PD-L1 status in lung adenocarcinoma better than in lung SCC. Model score was significantly predictive in the adenocarcinoma subtype (AUC ≈ 0.85, $P_{\text{permutation}} << 0.01$), but not in the SCC subtype (AUC = 0.64, $P_{\text{permutation}} = 0.18$). In SCC, the model score was higher for PD-L1+ samples than PD-L1− samples [Figure 3c and f], although the separation was not statistically significant ($P_{t\text{-test}} > 0.05$). The training cohort only contained 10 SCC cases (and 37 adenocarcinoma cases), which might been insufficient for the model to reliably identify subtle features associated with SCC PD-L1 status. A larger dataset with more SCC training examples could improve model performance.

The model performed well regardless of the specified PD-L1 positivity cutoff value. PD-L1 protein levels expressed by NSCLC tumors exist on a spectrum.[45] There is currently no consensus on a cutoff value to define PD-L1 positivity, resulting in a wide range of PD-L1 expression level cutoffs in clinical practice and trials.[15,17,46,47] For example, pembrolizumab was approved by the FDA as a single agent for metastatic NSCLC patients with tumor PD-L1 ≥1% who failed platinum-based chemotherapy.[15] However, PD-L1 expression cutoffs of ≥1%, ≥5%, and ≥10% were used in a Phase III clinical trial of NSCLC patients treated with nivolumab, which revealed that higher levels of PD-L1 expression are associated with greater efficacy of nivolumab.[13] Due to the discrepancy in these definitions of PD-L1 positivity, we conducted a sensitivity study to evaluate the impact of varying PD-L1 cutoffs on the model's ability to predict PD-L1 status. Adjusting the PD-L1 expression cutoffs from 5% to 50% had a moderate effect on the results with an AUC of 0.81 and 0.67, respectively. Despite this effect, a permutation test showed that all the predictions performed using various cutoffs were significant. The model recognizes complex H and E morphological features associated with different levels of tumor PD-L1 expression, suggesting it can be used in applications with varying PD-L1 cutoff values.

In the present study, tumor PD-L1 status was assessed using a Dako PD-L1 22c3 IHC assay, the FDA approved companion diagnostic for pembrolizumab prescription. Other assays, including Ventana SP142, Ventana SP263, and Dako 28-8, are used as diagnostic tests for atezolizumab, durvalumab, and nivolumab, respectively.[18,19,47] All IHC assays, except Ventana SP142, are concordant with the Dako 22c3 assay in assessing PD-L1 expression.[18,19,47] This suggests that our model can be used to predict PD-L1 status as determined by the Dako 28-8 and Ventana SP263 IHC assays; however, further studies are needed to investigate the model's performance in predicting Ventana SP142-assessed tumor PD-L1 status.

Our model predictions are also robust to simulated interpathologist variability. Assessments of tumor PD-L1

status can be affected by interpathologist disagreement.[47] Ratcliffe *et al*.[19] compared the agreement in NSCLC PD-L1 status assessment between a CLIA laboratory pathologist and an independent pathologist. The agreement between the two pathologists increased from moderate (75%) to excellent (96%) as the PD-L1 positivity cutoff increased from 1%–50%. To mimic the effect of interpathologist disagreement, we randomly shuffled 5%–30% of the PD-L1 labels in our test cohort. As expected, the power of our model in predicting PD-L1 status decreased as the proportion of shuffled labels increased [Figure 4b]. However, even in the case corresponding to maximum interpathologist divergence (25% shuffled samples), the model retained predictive power (AUC = 0.66, $P_{permutation} = 0.01$).

This study has two limitations. First, our training slides were annotated by only one pathologist. It is possible that the samples may include some bias arising from interpathologist variability. Despite this possibility, the model performed well in predicting PD-L1 status from H and E WSI. The model is expected to perform equally well or better with the use of consensus annotations from multiple pathologists. Second, the successful prediction of overall PD-L1 status did not require an accurate probability prediction at every individual location on each slide. For instance, in test cases with negative PD-L1 status, the average model score was approximately 5%, while the pathologist score was <1%, indicating that the model predicted some PD-L1− tumor regions as PD-L1+. Increasing the number of training examples will improve the model's ability to identify unique features that are associated with PD-L1+ tiles. This improvement is anticipated to allow more precise prediction of local PD-L1 expression.

Prospective studies with large, independent datasets and masked tumor PD-L1 status are needed to confirm our findings. Finally, future studies will need to investigate which interpretable morphological features contribute most to our deep learning model in order to extract biological meaning. For example, our model could be extended by adding the ability to classify PD-L1 expression among TILs. While overall tumor PD-L1+ status is commonly used as a response biomarker for checkpoint inhibitor immunotherapy,[7,15] PD-L1 expression on TILs is also strongly associated with improved response to checkpoint inhibitors in various cancer types.[48,49] A model that classifies tumor and lymphocyte PD-L1 status simultaneously would have broader applications for cancer patients. Furthermore, it would be an interesting future study to investigate if deep learning can uncover novel H and E and PD-L1 IHC features for enhanced prognosis and response prediction.

Our histopathological slides were prepared at a single CAP-accredited and CLIA-certified laboratory and were scanned using the same scanner. Other laboratories may follow different staining protocols, leading to color variations of the slides. Color normalization schemes, such as sparse nonnegative matrix factorization, have been used to standardized appearance of tissues slides and have shown to improve the performance of computer vision algorithms.[50,51] Thus, the robustness of our deep learning model can be further improved by implementing color normalization into our deep learning model. The diagnostic performance of scanners between different vendors has not been studied. The discordance and concordance in diagnosis between glass slides and WSI acquired from various scanners were about 15% and 85%, respectively.[52] Scanner variability is thus expected to have moderate effect on the performance of our deep learning model.

Conclusions

In this study, we developed a fast and robust deep learning model to predict tumor PD-L1 status from H and E WSIs in NSCLC. The prediction of tumor PD-L1 status was significant regardless of expression cutoff values, and the results were robust to interpathologist variability. This analysis may open new avenues to further developing an H and E image-based test to complement IHC staining for PD-L1 assessment, especially when there is insufficient tissue, and in some settings, a lack of resources for IHC staining.

Acknowledgments

The authors would like to thank Dr. Alexandria Bobe and Matthew Kase for editorial assistance. The authors would also like to express their appreciation to Erin McCarthy, Hunter Lane, Dr. Kevin White, Dr. Katie Igartua, Dr. Denise Lau, Dr. Martin Stumpe, and Dr. Rafi Pelossof for valuable comments on the manuscript.

Financial support and sponsorship

Nil.

Conflicts of interest

L.S., B.L.O., I.Y.H., C.W., N.B., B.M.M., T.J.T., and S.S.F.Y. are employees and/or shareholders of Tempus Labs. H.W. is an intern at Tempus Labs. T.L.T. was compensated by Tempus Labs for his participation as a pathologist.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394-424.
2. O'Rourke N, Macbeth F. Is concurrent chemoradiation the standard of care for locally advanced non-small cell lung cancer? A review of guidelines and evidence. Clin Oncol (R Coll Radiol) 2010;22:347-55.
3. Chang A. Chemotherapy, chemoresistance and the changing treatment landscape for NSCLC. Lung Cancer 2011;71:3-10.
4. Garg S, Gielda BT, Kiel K, Turian JV, Fidler MJ, Batus M, *et al.* Patterns of locoregional failure in stage III non-small cell lung cancer treated with definitive chemoradiation therapy. Pract Radiat Oncol 2014;4:342-8.
5. Teng MW, Ngiow SF, Ribas A, Smyth MJ. Classifying cancers based on T-cell infiltration and PD-L1. Cancer Res 2015;75:2139-45.
6. D'Incecco A, Andreozzi M, Ludovini V, Rossi E, Capodanno A, Landi L, *et al.* PD-1 and PD-L1 expression in molecularly selected non-small-cell lung cancer patients. Br J Cancer 2015;112:95-102.
7. Patel SP, Kurzrock R. PD-L1 expression as a predictive biomarker in cancer immunotherapy. Mol Cancer Ther 2015;14:847-56.

8. Kazandjian D, Suzman DL, Blumenthal G, Mushti S, He K, Libeg M, *et al.* FDA approval summary: Nivolumab for the treatment of metastatic non-small cell lung cancer with progression on or after platinum-based chemotherapy. Oncologist 2016;21:634-42.
9. Sundar R, Cho BC, Brahmer JR, Soo RA. Nivolumab in NSCLC: Latest evidence and clinical potential. Ther Adv Med Oncol 2015;7:85-96.
10. Sul J, Blumenthal GM, Jiang X, He K, Keegan P, Pazdur R, *et al.* FDA approval summary: Pembrolizumab for the treatment of patients with metastatic non-small cell lung cancer whose tumors express programmed death-ligand 1. Oncologist 2016;21:643-50.
11. Brahmer JR, Tykodi SS, Chow LQ, Hwu WJ, Topalian SL, Hwu P, *et al.* Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. N Engl J Med 2012;366:2455-65.
12. Meng X, Huang Z, Teng F, Xing L, Yu J. Predictive biomarkers in PD-1/PD-L1 checkpoint blockade immunotherapy. Cancer Treat Rev 2015;41:868-76.
13. Borghaei H, Paz-Ares L, Horn L, Spigel DR, Steins M, Ready NE, *et al.* Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med 2015;373:1627-39.
14. Abdel-Rahman O. Correlation between PD-L1 expression and outcome of NSCLC patients treated with anti-PD-1/PD-L1 agents: A meta-analysis. Crit Rev Oncol Hematol 2016;101:75-85.
15. U.S. Food & Drug Administration (FDA). KEYTRUDA Label. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/label/2017/125514s024lbl.pdf#page=46. [Last accessed on 2019 Feb 12].
16. Baas P, Garon EB, Herbst RS, Felip E, Perez-Gracia JL, Han JY, *et al*. Relationship between level of PD-L1 expression and outcomes in the KEYNOTE-010 study of pembrolizumab vs. docetaxel for previously treated, PD-L1-Positive NSCLC. J Clin Orthod 2016;34:9015.
17. Roach C, Zhang N, Corigliano E, Jansson M, Toland G, Ponto G, *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;24:392-7.
18. Rimm DL, Han G, Taube JM, Yi ES, Bridge JA, Flieder DB, *et al.* A prospective, multi-institutional, pathologist-based assessment of 4 immunohistochemistry assays for PD-L1 expression in non-small cell lung cancer. JAMA Oncol 2017;3:1051-8.
19. Ratcliffe MJ, Sharpe A, Midha A, Barker C, Scott M, Scorer P, *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;23:3585-91.
20. Patel K, Strother RM, Ndiangui F, Chumba D, Jacobson W, Dodson C, *et al.* Development of immunohistochemistry services for cancer care in Western Kenya: Implications for low-and middle-income countries. Afr J Lab Med 2016;5:187.
21. Adeyi OA. Pathology services in developing countries-the West African experience. Arch Pathol Lab Med 2011;135:183-6.
22. Cardiff RD, Miller CH, Munn RJ. Manual hematoxylin and eosin staining of mouse tissue sections. Cold Spring Harb Protoc 2014;2014:655-8.
23. Feldman AT, Wolfe D. Tissue processing and hematoxylin and eosin staining. Methods Mol Biol 2014;1180:31-43.
24. Velcheti V, Schalper KA, Carvajal DE, Anagnostou VK, Syrigos KN, Sznol M, *et al.* Programmed death ligand-1 expression in non-small cell lung cancer. Lab Invest 2014;94:107-16.
25. McLaughlin J, Han G, Schalper KA, Carvajal-Hausdorf D, Pelekanou V, Rehman J, *et al.* Quantitative assessment of the heterogeneity of PD-L1 expression in non-small-cell lung cancer. JAMA Oncol 2016;2:46-54.
26. Denkert C, Wienert S, Poterie A, Loibl S, Budczies J, Badve S, *et al.* Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: Results of the ring studies of the international immuno-oncology biomarker working group. Mod Pathol 2016;29:1155-64.
27. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, *et al.* Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 2017;318:2199-210.
28. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.
29. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, *et al.* A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
30. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Velázquez Vega JE, *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. Proc Natl Acad Sci U S A 2018;115:E2970-9.
31. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018;24:1559-67.
32. Gertych A, Swiderska-Chadaj Z, Ma Z, Ing N, Markiewicz T, Cierniak S, *et al.* Convolutional neural networks can accurately distinguish four histologic growth patterns of lung adenocarcinoma in digital slides. Sci Rep 2019;9:1483.
33. Schaumberg AJ, Rubin MA, Fuchs TJ. H and E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer. bioRxiv; 2018. p. 064279.
34. Couture HD, Williams LA, Geradts J, Nyante SJ, Butler EN, Marron JS, *et al.* Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. NPJ Breast Cancer 2018;4:30.
35. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw 2015;61:85-117.
36. U.S. Food and Drug Administration (FDA). PD-L1 IHC 22C3 pharmDx. Summary of Safety and Effectiveness Data. Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf15/P150013S006b.pdf. [Last accessed on 2019 Feb 12].
37. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, *et al.* QuPath: Open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
38. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition; 2015. Available from: http://arxiv.org/abs/1512.03385. [Last accessed 2019 Jun 21].
39. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv [cs.CV]; 2014. Available from: http://arxiv.org/abs/1409.1556. [Last accessed 2019 Jun 21].
40. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, *et al.* ImageNet Large Scale Visual Recognition Challenge. arXiv [cs.CV]; 2014. Available from: http://arxiv.org/abs/1409.0575. [Last accessed 2019 Jun 21].
41. Ioffe S, Szegedy C. Batch Normalization: Accelerating Dee p Network Training by Reducing Internal Covariate Shift. arXiv [cs.LG]; 2015. Available from: http://arxiv.org/abs/1502.03167. [Last accessed 2019 Jun 21].
42. Wimberly H, Brown JR, Schalper K, Haack H, Silver MR, Nixon C, *et al.* PD-L1 expression correlates with tumor-infiltrating lymphocytes and response to neoadjuvant chemotherapy in breast cancer. Cancer Immunol Res 2015;3:326-32.
43. Kitano A, Ono M, Yoshida M, Noguchi E, Shimomura A, Shimoi T, *et al.* Tumour-infiltrating lymphocytes are correlated with higher expression levels of PD-1 and PD-L1 in early breast cancer. ESMO Open 2017;2:e000150.
44. Vassilakopoulou M, Avgeris M, Velcheti V, Kotoula V, Rampias T, Chatzopoulos K, *et al.* Evaluation of PD-L1 expression and associated tumor-infiltrating lymphocytes in laryngeal squamous cell carcinoma. Clin Cancer Res 2016;22:704-13.
45. Kerr KM, Hirsch FR. Programmed death ligand-1 immunohistochemistry: Friend or foe? Arch Pathol Lab Med 2016;140:326-31.
46. Herbst RS, Baas P, Kim DW, Felip E, Pérez-Gracia JL, Han JY, *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;387:1540-50.
47. Büttner R, Gosney JR, Skov BG, Adam J, Motoi N, Bloom KJ, *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;35:3867-76.
48. Janzic U, Kern I, Janzic A, Cavka L, Cufer T. PD-L1 expression in squamous-cell carcinoma and adenocarcinoma of the lung. Radiol Oncol 2017;51:357-62.
49. Herbst RS, Soria JC, Kowanetz M, Fine GD, Hamid O, Gordon MS,

et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature 2014;515:563-7.
50. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, *et al.* Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans Med Imaging 2016;35:1962-71.
51. Bejnordi BE, Litjens G, Timofeeva N, Otte-Höller I, Homeyer A, Karssemeijer N, *et al.* Stain specific standardization of whole-slide histopathological images. IEEE Trans Med Imaging 2016;35:404-15.
52. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, *et al.* Validating whole slide imaging for diagnostic purposes in pathology: Guideline from the college of American Pathologists Pathology and Laboratory Quality Center. Arch Pathol Lab Med 2013;137:1710-22.