## Title

A Pipeline for Evaluation of Machine Learning/Artificial Intelligence Models to Quantify Programmed Death Ligand 1 Immunohistochemistry



## Journal

Laboratory Investigation, 104(6)

## ISSN

0023-6837

## Authors

Knudsen, Beatrice S
Jadhav, Alok
Perry, Lindsey J
et al.

## Publication Date

2024-06-01

## DOI

10.1016/j.labinv.2024.102070

Lab Invest 104 (2024) 102070

LABORATORY INVESTIGATION

journal homepage: https://laboratoryinvestigation.org/

Research Article

# A Pipeline for Evaluation of Machine Learning/Artificial Intelligence Models to Quantify Programmed Death Ligand 1 Immunohistochemistry

Beatrice S. Knudsen[a,b,*], Alok Jadhav[c], Lindsey J. Perry[a], Jeppe Thagaard[d], Georgios Deftereos[e], Jian Ying[f], Ben J. Brintz[f], Wei Zhang[a,b,*]

[a] *Department of Pathology, University of Utah, Salt Lake City, Utah;* [b] *Huntsman Cancer Institute, University of Utah, Salt Lake City, Utah;* [c] *ARUP, Salt Lake City, Utah;* [d] *Visiopharm A/S, Hoersholm, Denmark;* [e] *Department of Pathology, UCSF, San Francisco, California;* [f] *Department of Internal Medicine, University of Utah, Salt Lake City, Utah*

ARTICLE INFO

*Article history:*
Received 5 October 2023
Revised 8 April 2024
Accepted 18 April 2024
Available online xxx

*Keywords:*
cancer segmentation
digital pathology
programmed death ligand 1
tumor proportion scores

ABSTRACT

Immunohistochemistry (IHC) is used to guide treatment decisions in multiple cancer types. For treatment with checkpoint inhibitors, programmed death ligand 1 (PD-L1) IHC is used as a companion diagnostic. However, the scoring of PD-L1 is complicated by its expression in cancer and immune cells. Separation of cancer and noncancer regions is needed to calculate tumor proportion scores (TPS) of PD-L1, which is based on the percentage of PD-L1-positive cancer cells. Evaluation of PD-L1 expression requires highly experienced pathologists and is often challenging and time-consuming. Here, we used a multi-institutional cohort of 77 lung cancer cases stained centrally with the PD-L1 22C3 clone. We developed a 4-step pipeline for measuring TPS that includes the coregistration of hematoxylin and eosin, PD-L1, and negative control (NC) digital slides for exclusion of necrosis, segmentation of cancer regions, and quantification of PD-L1+ cells. As cancer segmentation is a challenging step for TPS generation, we trained DeepLab V3 in the Visiopharm software package to outline cancer regions in PD-L1 and NC images and evaluated the model performance by mean intersection over union (mIoU) against manual outlines. Only 14 cases were required to accomplish a mIoU of 0.82 for cancer segmentation in hematoxylin-stained NC cases. For PD-L1-stained slides, a model trained on PD-L1 tiles augmented by registered NC tiles achieved a mIoU of 0.79. In segmented cancer regions from whole slide images, the digital TPS achieved an accuracy of 75% against the manual TPS scores from the pathology report. Major reasons for algorithmic inaccuracies include the inclusion of immune cells in cancer outlines and poor nuclear segmentation of cancer cells. Our transparent and stepwise approach and performance metrics can be applied to any IHC assay to provide pathologists with important insights on when to apply and how to evaluate commercial automated IHC scoring systems.



## Introduction

Immunohistochemistry (IHC) is used to improve diagnostic accuracy and resolve specific prognostic and treatment-related questions.[1] Glass slides are stained with antibodies that bind to proteins in all cell types on the slide. In cancer tissues, the IHC-

* Corresponding authors.
E-mail addresses: Wei.zhang@hci.utah.edu (W. Zhang), beatrice.knudsen@path.utah.edu (B.S. Knudsen).

0023-6837/$ - see front matter 
https://doi.org/10.1016/j.labinv.2024.102070

targeted proteins are often expressed in benign cells in the tumor microenvironment in addition to cancer cells. These benign cells can be mistakenly included in the assessment of cancer regions and increase false-positive rates in IHC staining results. To mitigate this problem, pathologists must visually separate cancer and noncancer cells when reporting IHC results. However, when software quantitates IHC, it is not always apparent whether and how models distinguish IHC signals in cancer versus noncancer cells. This issue is particularly serious in the context of programmed death ligand 1 (PD-L1) staining.[2]

The cell surface protein, PD-L1, is an inhibitor of T-cell cytotoxicity and a target for immunotherapy.[3] Expression of PD-L1 is usurped by cancer cells to evade the immune response.[4] Under the control of immunoregulatory cytokines, PD-L1 is mostly expressed in cancer cells and macrophages. Multiple Food and Drug Administration–approved or College of American Pathologists/Clinical Laboratory Improvement Amendments (CAP/CLIA) (laboratory developed test [LDT])-approved tissue staining assays employ 22C3, 28-8, SP142, SP265, and 73-10 antibodies to visualize PD-L1 expression as a companion diagnostic for treatment with checkpoint inhibitors (pembrolizumab, nivolumab, atezolizumab, duravalumab, and avelumab). Treatment decisions are made in part on the tumor proportion score (TPS—percentage of positive and viable tumor cells) or combined positive score (CPS—ratio of positive, viable cancer and immune cells to viable cancer cells).[2]

The PD-L1 IHC staining can be difficult to quantify through a microscope. Pathologists benefit from additional training to improve the accuracy and reproducibility of manual PD-L1 scoring, and challenging cases require a consensus of multiple pathologists.[5] To assist with the evaluation of PD-L1 expression, multiple teams working in the field of pathology image analysis developed algorithms for PD-L1 quantification.[6-9] Machine learning models, developed with digital slides from different cancer types and antibody assays, work well when comparing the concordance of computer-generated TPS with manual TPS (mTPS). However, the performance of the models to distinguish PD-L1 expression in cancer and immune cells at a cell level is lacking. To gain the confidence of pathologists in the ability of models to output accurate TPS or CPS scores, cancer cells and immune cells need to be separated and individually analyzed. This cancer segmentation step is challenging because it needs to be performed in the absence of the Eosin stain in the slide. Cancer cells may lack PD-L1 expression, leaving the hematoxylin channel as the only source of data for the detection of cancer cells.

Pathologists follow a consistent and systematic workflow for generating PD-L1 expression scores. They first identify the tumor regions and then the areas of necrosis and the extent of immune cell infiltration. To determine the TPS in a slide, pathologists evaluate the percentage of tumor cells with PD-L1 membrane staining. We propose a pipeline for a machine learning model that generates TPS for PD-L1-stained digital slides and that is inspired by the workflow used by pathologists. To facilitate clinical adoption and interoperability with slide management systems, we utilized the Visiopharm software package for the configuration of a 4-step pipeline. Due to the importance of cancer segmentation accuracy in the overall performance of PD-L1 quantification, we experimentally examined multiple variables that affect the performance of the cancer segmentation step. Except for the last step, which is a rule-based code calibrated on PD-L1 cell surface expression, our stepwise approach can be generalized broadly to evaluate and optimize automated solutions for IHC quantification.

## Methods

### *Cases, Programmed Death Ligand 1 Staining, Slide Scanning, and Manual Annotations*

A total of 77 lung cancer cases were included in this study (Fig. 1A). All cases were stained in the CLIA/CAP laboratory using the Food and Drug Administration–approved assay for staining with the anti-PD-L1 antibody 22C3.[10] In addition, an hematoxylin and eosin (H&E)-stained and a negative antibody control (NC) slide were available from the same tissue block as the PD-L1 slide. Cases with tissue blocks from surgical resections or from large core biopsies with a diagnosis of non–small cell carcinoma of the lung, including both adenocarcinoma and squamous carcinomas, were enrolled for analysis. Slides from endobronchial biopsies and fine needle aspiration were excluded. Slides were scanned on the Aperio AT2 slide scanner (Leica Biosystems, Inc). This study was conducted under the oversight of IRB 00091019.

Manual annotations for training, validation, and testing were generated in all digital PD-L1 and NC slides, whereas the H&E slide was used as a reference for the identification of cancer regions. Algorithms were trained using PD-L1 or NC slides separately or together. Cancer segmentation results were evaluated by

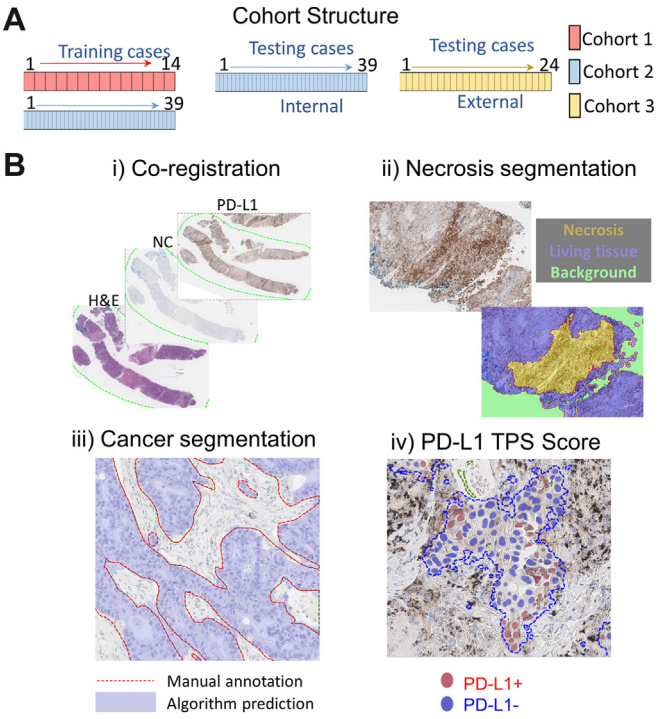


**Figure 1.**
(A and B) Project workflow. (A) Schematic of model training and performance testing for cancer segmentation. Programmed death ligand 1 (PD-L1) training and testing cases. PD-L1 and corresponding negative control patches from cases in cohorts 1 ($n$ = 14) and 2 ($n$ = 39) were used for training of cancer segmentation algorithms in slides stained by PD-L1 immunohistochemistry. Cohort 3 ($n$ = 24) was reserved for external testing. (B) Image analysis pipeline. (i) Image coregistration. hematoxylin and eosin, negative control, and PD-L1 whole slide images were coregistered for each case. (ii) Elimination of areas with necrosis. A U-Net model was trained on PD-L1 immunohistochemistry cases to outline necrotic cancer regions. (iii) Cancer segmentation. DeepLab V3 was trained to identify cancer regions. (iv) Percentage of PD-L1-positive cancer cells (tumor proportion scores). Viable cells in cancer regions are classified into PD-L1-positive and -negative to calculate the tumor proportion score, that is, the percentage of PD-L1-positive, viable cancer cells.

calculating the intersection over union (IoU) of computer-generated versus manual cancer outlines. Digital TPS (dTPS) was compared with the mTPS score using the percent accuracy as a metric. The mTPS was abstracted from the pathology report used as the reference standard.

Cases are divided into 3 separate cohorts for model training, internal testing, and external testing (Fig. 1A). Cohort 1 consists of 14 cases (training set), cohort 2 of 39 cases (training set or internal test set), and cohort 3 of 24 cases (held-out test set). Two pathologists (B.S.K. and L.J.P.) annotated the images. In cohort 1, the entire cancer region was annotated in the NC whole slide image (WSI), whereas in the PD-L1 slide, 2 patches were annotated in the cancer region. In cohort 2, we annotated 2 patches after registration of NC and PD-L1 for training and 1 patch for testing (Supplementary Fig. S1). In cohort 3, we annotated 1 patch in the registered PD-L1 and NC for testing. All patches possess the same size and are obtained after coregistration of NC and PD-L1 WSIs, providing samples of the same tumor region from 2 parallel slides of the tissue block.

### *Visiopharm Software Application*

All scanned WSIs are analyzed using commercial Visiopharm software, which contains several pull-down options for the application of deep convolutional neural networks (CNN) (Fig. 1B). The analysis pipeline to calculate PD-L1 TPSs consists of 4 modules within Visiopharm. These include image coregistration (rule-based), necrosis removal (CNN), cancer segmentation (CNN), and PD-L1 scoring (rule-based).

### *Image Registration*

For registration, we uploaded WSIs of PD-L1 IHC, NC, and H&E from the same case to the Visiopharm coregistration module called “Tissuealign” and digitally aligned 2 slides using an affine registration algorithm (Tissuealign, Visiopharm A/S).

### *Necrosis Segmentation and Elimination of Necrotic Regions From the Images*

We trained a U-Net model to segment necrotic tissue regions in PD-L1 WSI. Necrotic areas were identified in the H&E images, and the outlines of necrosis were used to mark necrotic regions in the registered PD-L1 IHC-WSI. Tiles of 256 × 256 pixels at 10× were extracted within the annotated polygons from the 14 slides in cohort 1 and used to train the U-Net model. The performance of the model was visually evaluated.

### *Cancer Segmentation*

We trained DeepLabV3 provided by Visiopharm for tumor segmentation using tiles of 256 × 256 pixels at 10× magnification. The input into the model consisted of NC, PD-L1, or NC plus PD-L1 tiles. The cancer segmentation algorithms were trained using a learning rate of 1.0e to 0.5 and applying a median filter for smoothing the image features in the first layer. We used the IoU with regard to the manual annotations to determine the performance of the model.

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ Union}$$

An IoU = 1 indicates a perfect overlap of manual and algorithmic cancer outlines, whereas an IoU = 0 indicates that there is no overlap. The closer the IoU is to 1, the better the performance of the model. To determine the heterogeneity of TPS in a subset of 14 cases in cohort 1, the cancer region was divided into 5 regions of similar size, and separate TPS scores were obtained from each subregion.

### *Programmed Death Ligand 1 Quantification and Tumor Proportion Scores*

The first step in the assessment of PD-L1 expression in cancer cells involved a rule-based nuclear segmentation within segmented cancer regions. The nuclear segmentation uses the hematoxylin channel and relies on image-processing methods such as blob detection, median filtering, thresholding, and spatial domain linear filtering. After removing the background noise with filters that encompass parameters of cell morphology, nuclear outlines are dilated with a 3 × 3 kernel and eroded until a virtual cell outline is generated. The cell outlines are overlaid with a PD-L1 mask in the deconvoluted diaminobenzidine channel, and positive PD-L1 pixels are determined for each cell. A threshold is set visually for the number of PD-L1-positive pixels inside the dilated cell outline to call a cell PD-L1-positive. PD-L1-positive cells are marked by red nuclei, and PD-L1-negative cells are marked by blue nuclei in Supplementary Figure S2. The threshold numbers vary between laboratories performing PD-L1 staining. Therefore, a technical expert from the Visiopharm support team helped determine the optimized thresholds for quantifying positive PD-L1 cells for the cases in this study. TPS is calculated by (PD-L1+ Tumor Cells)/(Total Viable Tumor cells) *100 in the WSI or for specific regions of interest.

### *Statistical Analysis*

IoU was used as the performance metric to evaluate each cancer segmentation algorithm. Violin plots were used to visualize differences in IoU between algorithms. Within each violin plot, the mean value is indicated by a white line. We also employed box plots, in which the median value is represented by a white line, and the edges of the box encompass the interquartile range (25th-75th percentiles) of the data points. Heatmaps were used to illustrate the concordance of dTPS and mTPS. The accuracy of the dTPS was calculated based on the mTPS cutoffs of <1%, 1% to 50%, and >50%. We used OriginLab Pro 2022b to generate all visualizations.

To determine the significance of the difference in IoUs between the 2 models, we calculated the mean of the paired differences in IoUs and used a permutation test. Specifically, in each case, we created a permutation by randomizing the membership within each pair and calculated the mean of the paired differences using the permuted data. We generated 1,000 permutations (D1, …, D1000) to approximate the null distribution of the mean of the paired difference of IoUs. We also calculated the *P* values based on the 2-tailed permutation test using the following formula:

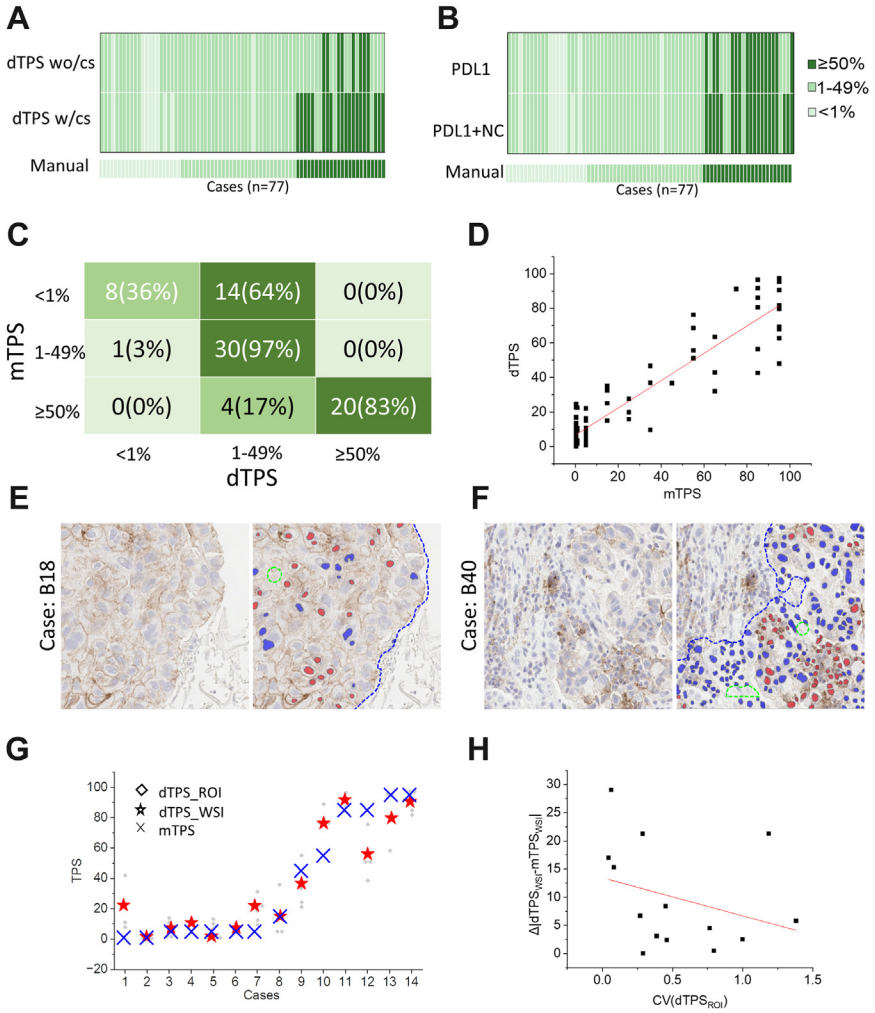


**Figure 2.**
Comparison of digital tumor proportion scores (dTPS) and manual tumor proportion scores (mTPS) tumor proportion scores. (A) Comparison of dTPS with cancer segmentation (dTPS/w cs) and without (dTPS/wo cs) cancer segmentation. TPS scores are obtained from glass slides (mTPS) and whole slide images (WSIs) (dTPS). (B) Case-wise comparison of mTPS to dTPS generated with cancer segmentation by models trained on programmed death ligand 1 (PD-L1) or on PD-L1 plus negative control (NC) patches from 53 cases for cancer segmentation. TPS scores are divided into 3 clinically relevant classes (<1%, 1%-49%, and ≥50%). The entire tumor region segmented by the model is used for calculation of TPS. All the cases ($n = 77$) are included for TPS measurements after cancer segmentation. (C) Confusion matrix showing the concordance between mTPS and dTPS scores. The PDL1 + NC cancer segmentation model was applied to generate the cancer outlines. (D) Scatterplot of mTPS and dTPS ($n = 77$). The Spearman correlation coefficient = 0.84 between dTPS and mTPS. (E and F) Representative cases with a discordance between dTPS and mTPS. Blue lines represent tumor outlines that are generated by the PD-L1 + NC model. Green outlines indicate the background pixels inside the cancer region. Solid red nuclei illustrate PD-L1-positive cancer cells, and blue nuclei mark PD-L1-negative cancer cells. (G) Effect of tumor heterogeneity on TPS score. Tumor regions in each case were divided into 5 regions of interest (ROIs) of approximately the same area. dTPS scores were obtained from each ROI. dTPS was also obtained from the WSI. The mTPS is from the pathology report. (H) Relationship between discrepancy between mTPS and dTPS in the WSI and tumor heterogeneity as quantified by the coefficient of variation of dTPS measurements across the 5 ROIs from each case ($n = 14$).

$\frac{\Sigma_{i=1}^{1000} I(|D_0| > |D_i|)}{1000}$, using 0.05 as a threshold for statistical significance.

The accuracy of the dTPS score using the mTPS score as the truth was calculated as the ratio of the number of correctly classified cases to the total of cases under evaluation. Furthermore, we used the 2-way mixed effects intraclass correlation coefficient (ICC) to assess reliability of the dTPS. Reliability refers to the extent to which the digital scores replicate the mTPS scores.[11] We opted for the 2-way mixed effects models as they account for correlation within a single scored case, considering only the "raters"—the digital tool and manual call—as the raters of interest. We provide the estimate and 95% confidence interval (CI) for the ICC.

## Results

We determine whether cancer segmentation improves the accuracy of PD-L1 TPS scores by comparing TPS scores with and without cancer segmentation, using the mTPS score from the pathology report as the gold standard. After grouping cases into 3 groups that are defined by TPS < 1%, 1% ≤ TPS < 50%, and TPS > 50%, we observed greater accuracy (75% vs 57%) when using the cancer segmentation module in the pipeline (Fig. 2A). As a next step, we compared the training of the cancer segmentation algorithm with patches from PD-L1 stained digital slides versus PD-L1 plus NC patches for data augmentation. The TPS results were more accurate (75% vs 68%) using the cancer segmentation algorithm

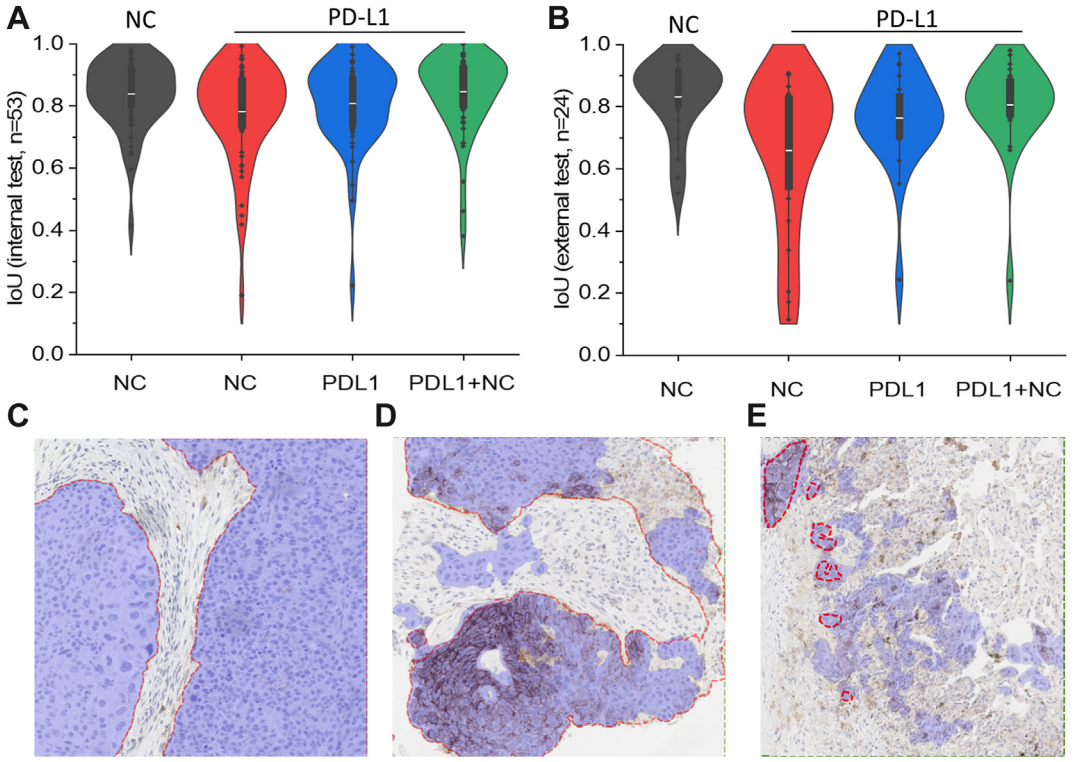


**Figure 3.**
Performance testing for cancer segmentation in programmed death ligand 1 (PD-L1) stained digital slides. (A and B) Performance evaluation in internal test patches (A) and external test cases (B). Four models, depicted on the x-axis, were tested on held-out patches, not used for training, from cohorts 1 and 2. Models were either trained on negative control (NC) patches, PD-L1 patches, or on a combination of NC and PD-L1 patches. NC and PD-L1 patches from each case are coregistered and represent the same tumor area. Labels above the violins indicate whether NC or PD-L1 patches were used for testing. NC cases are used as the baseline (black violin). (C–E) Tumor outlines in PD-L1-stained patches. (C) Representative case with high agreement (intersection over union [IoU] = 0.99) between manual (red line) and predicted (blue mask) cancer region. (D) Representative case with medium agreement (IoU = 0.82) between manual and predicted cancer region. (E) Representative case with low agreement (IoU = 0.24) between manual and predicted cancer regions.

trained on NC plus PD-L1 patches compared with the segmentation algorithm trained only on PD-L1 patches (Fig. 2B).

After calculating the percentage of PD-L1-positive cells in segmented cancer regions, we determined the accuracy of dTPS scores using mTPS scores as the gold standard. As shown in the confusion matrix in Figure 2C, there were accuracies of 97% and 83% in the TPS >1% to 49% and TPS ≥ 50% groups. However, 64% of cases with mTPS < 1% were scored as > 1% by the dTPS pipeline. The overall Spearman rank correlation for 3 categorial groups between WSI dTPS and mTPS from the pathology report is 0.84 (95% CI: 0.74-0.90) (Fig. 2D). In addition, a 2-way mixed effects ICC was used to measure agreement between the manual and digital scores, showing the agreement at 0.914 (95% CI: 0.867-0.944).

When examining the reasons for discrepancies between dTPS and mTPS, we identified the following 2 major problems: (1) incomplete nuclear identification and generation of cell boundaries (Fig. 2E) and (2) inclusion of immune cells in cancer outlines (Fig. 2F). Case B18 is an example of poor nuclear segmentation because DNA, in this case, was damaged during tissue processing leading to weak hematoxylin staining. Despite optimized settings of the rule-based nuclear segmentation code in Visiopharm to improve the nuclear segmentation results (see methods), the B18 case revealed poor nuclear detection results. Another example, case B40, illustrates the problem generated by inclusion of PD-L1-positive immune cells in the cancer region, increasing the dTPS compared with mTPS score.

PD-L1 staining can be affected by tumor heterogeneity and heterogeneity of immune cell infiltration. To determine whether tumor heterogeneity generates a bias in manual PD-L1 scoring by pathologists, we compared the manual score of the WSI with 5 tumor regions of approximately equal size that were analyzed by the digital pipeline. The results are shown in Figure 2G. Except for cases 10 and 12, there is good agreement between manual and digital PD-L1 assessment, suggesting that tumor heterogeneity of PD-L1 expression does not introduce a bias in manual scoring. For a quantitative evaluation of tumor heterogeneity, we calculated the coefficient of variation (%CV) by dividing the standard deviation of dTPS from the 5 regions of interest by the mean TPS score. The %CV is plotted against the absolute difference between mTPS and dTPS (Fig. 2H). Plotting the %CV against the difference in dTPS and mTPS scores demonstrates an inverse correlation between the tumor heterogeneity, ie, %CV, and the difference between dTPS and mTPS scores. The negative slope indicates that the discrepancy between digital and manual PD-L1 scores declines as PD-L1 staining heterogeneity increases. From this result, we conclude that the heterogeneity of PD-L1 expression in cancer does not introduce a bias in manual PD-L1 TPS scoring in the cases from cohort 1.

Next, we further explored the cancer segmentation task through a series of ablation studies. We questioned how well the model trained on NC hematoxylin-stained digital slides generalizes to patches from PD-L1-stained slides (Supplementary Fig. S3). We employ the IoU between computer-generated and manual outlines as the performance metric. Training on NC patches and testing on NC patches is used as a reference (black violins in Fig. 3A, B) with an average IoU of 0.84 (internal test set in Fig. 3A) and 0.82 (external test set in Fig. 3B). When we applied the NC

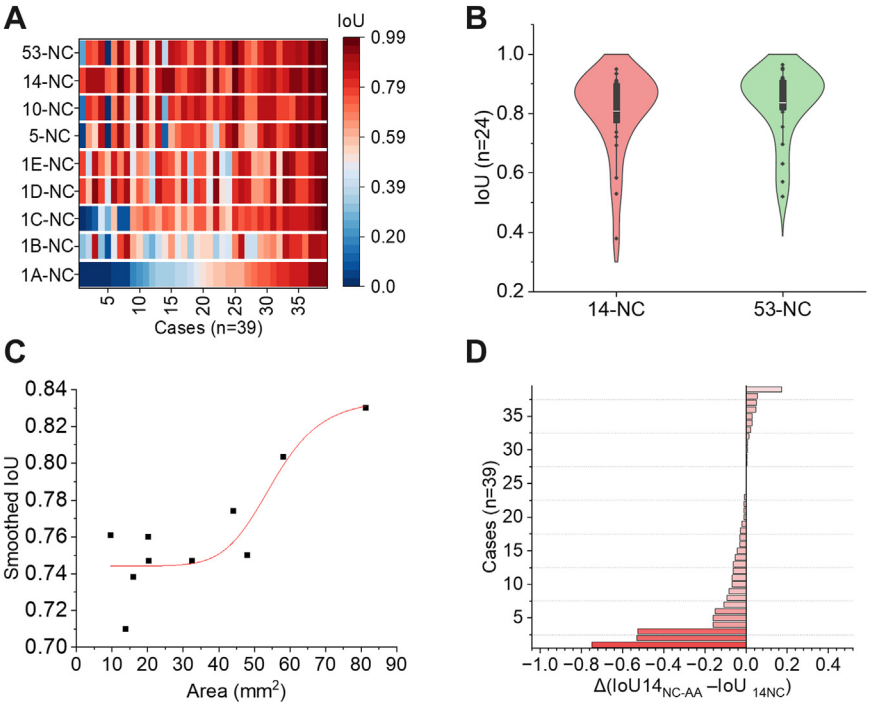


**Figure 4.**

Cancer segmentation in the hematoxylin channel. (A) Effect of training case numbers on algorithm performance. A heatmap shows the intersection over union (IoU) for each case from cohort 2 in a separate column ($n = 39$). Cases used for algorithm training are indicated on the y-axis. (B) Comparison of IoU in testing data set using algorithms trained on 14 and 53 cases on external cases in cohort 3. (C and D) Effect of cancer area used for training on model performance. (C) Effect of cancer area size on IoU. Nonlinear fitting of smoothed IoU to the cancer area that the algorithm is trained on. (D) Model performance in individual cases. Case-wise differences in performance between algorithms trained on the entire cancer region from each case in cohort 1 versus the cancer region after size adjustment. IoUs were calculated using manual outlines as the reference.

model to internal or external PD-L1 test patches that represent the same tumor regions as in NC patches, the average IoU decreased significantly (red violins). Therefore, we proceeded to train a cancer segmentation model directly on the unmixed hematoxylin channel of PD-L1 patches. Models trained on the hematoxylin image from PD-L1 patches (blue violins) possess a greater average IoU compared with models trained on NC patches. The best performance (mIoU = 0.79 on external data) is obtained when models are trained on coregistered NC and PD-L1 patches, demonstrating that augmentation of data through inclusion of NC patches improves cancer segmentation accuracy in PD-L1 WSIs.

When inspecting why models fail (Fig. 3C-E), we observe both false-positive stromal segmentation and lack of tumor segmentation (Fig. 3D). In addition, we find that the models struggle with immune cell regions, despite their training on 5 cases of NC lymph nodes (negative for carcinoma). In summary, the most accurate cancer segmentation in PD-L1 patches is obtained through augmenting PD-L1 patches with patches from NC WSI from the same cases, which are generated during the clinical tissue staining process.

Finally, we systematically determined the amount of data that is required to optimize the cancer segmentation task when only using the hematoxylin channel. In the first experiment, Deeplab V3 was trained using only 1 case, and the training was repeated 5 times with a different case from cohort 1. The trained models (1A—E-NC) were applied to the same 39 manually annotated patches from cohort 2 (not used for training). The results of IoUs for each case are shown as a heatmap in Figure 4A and reveal significant differences (analysis of variance, $P < .001$) in model performance depending on which case was used for training. Next, we used 3 different batches of 5 cases from cohort 1 for training and observed consistent mIoUs on the 39 cases from cohort 2 (Supplementary Fig. S4). Then, we compared models trained on 5, 10, or 14 cases from cohort 1. The IoU improved with increasing numbers of cases used for training (Fig. 4A). Finally, we compared the model trained on 14 cases from cohort 1 with a model trained on 53 cases from cohorts 1 and 2. When tested on the 24 held-out cases from cohort 3 (Fig. 4B), no significant difference was observed in the performance of the 2 models ($P = .29$). However, the model trained on 53 cases generalized better as observed by the decrease in the variance of IoUs among the 24 cases.

Next, we addressed whether the amount of data from each case affected the performance of the algorithm. When we compared the cancer areas in the 14 cases from cohort 1, 2 cases had larger cancer areas (Supplementary Fig. 5). When we reduced the cancer area in the 2 cases and repeated the training of the 14-negative control adjusted area model. The mIoU of this model significantly decreased compared with the original 14-NC model (examples of cancer tissue outline in Supplementary Fig. S6). This alerted us to the possibility that the cancer area per case and not just the number of cases mattered for training. To evaluate the role of the cancer regions in a more systematic fashion, we plotted the smoothened IoU against the cancer region used for training (Fig. 4C). The resulting curve reached a plateau (80 $mm^2$), which corresponded to a cancer area less than the sum of the cancer areas of the 14 cases in cohort 1. Finally, we determined, using cases from cohort 2, whether the decrease in performance between the 14-NC and 14-negative control adjusted area models is case-specific or occurs equally across all the cases. The greatest reduction in IoU amounted to a difference in the IoU of 0.78. However, 6 cases experienced a small increased IoU with the 14-negative control adjusted area compared with the 14-NC model (Fig. 4D).

## Discussion

In our study, we illustrate a strong agreement between manual and digital TPS, yet highlight the necessity for improvement, particularly at the 1% cutoff of TPS, crucial for determining the suitability of treatment with immune checkpoint inhibition. Furthermore, we demonstrate that dTPS is more accurate if cancer regions are outlined prior to measuring TPS. Finally, we provide preliminary evidence that heterogeneity of PD-L1 expression in tumor and immune cell regions does not introduce a significant bias in mTPS in our practice of fellowship-trained, subspecialized molecular pathologists. The augment experiments demonstrate that the performance of outlining cancer in the hematoxylin channel depends on the number of cases as well as the area of cancer used for training. The results from an unbiased consecutive number of lung cancer cases demonstrate that the training of the DeepLab V3 model can be accomplished with as few as 15 cases and does not improve significantly when more cases are used. The machine-assisted scoring of IHC increases the accuracy and reproducibility of manual scoring and reduces the burden on pathologists. However, algorithmic scoring systems that are trained end-to-end to output percentages of positive cells, while providing good performance, are plagued by a lack of transparency. Our study provides the following 3 novel insights that might increase transparency and help with clinical adoption of IHC scoring algorithms: (1) we explicitly demonstrate the benefits of cancer segmentation to increase the accuracy of measuring IHC-positive cancer cells in WSIs; (2) our results show that the performance of the cancer segmentation algorithm can be improved through augmentation of tiles from the IHC-WSI by NC tiles; and (3) we also demonstrate that a surprisingly small number of cases are needed to adapt a pretrained model to the unique characteristics of cases of at our local site.

Although not a perfect biomarker, PD-L1 serves as the only tissue marker, to date, to guide treatment decisions with immune checkpoint inhibitors (ICIs). In lung cancer, 3 treatment groups exist. Cases with less than 1% of PD-L1-positive cancer cells are in general not treated with ICIs. These cases are easy to identify because they have no or weakly detectable PD-L1 staining. Artificial intelligence models, in general, are more sensitive than pathologists to identify PD-L1-positive cells. Therefore, using retrospective clinical trial cases, Baxi et al[6] proposed a 5% cutoff for digital TPS to distinguish PD-L1-negative and -positive cases instead of the 1% mTPS cutoff. On the other side of the spectrum, cases with $\geq$50% PD-L1-positive cancer cells who may receive single-agent ICI treatment frequently express PD-L1 in cancer and immune cells. These cases are easy to identify because they exhibit many PD-L1-positive cells with dark membrane staining in all viable tumor regions, and an exact separation between cancer and noncancer cells is not as critical. The main group benefiting from computer-assisted scoring is the group with 1% to 49% PD-L1-positive cancer cells. Lung cancer patients in this group are treated with ICI plus chemotherapy. This intermediate group possesses the greatest heterogeneity in cancer and immune cell staining. The separation of cancer and immune cells can be learned by an algorithm to help pathologists with challenging cases.

Several algorithms[12-16] have been published that use machine learning or deep learning models to calculate TPS in an automated fashion and by applying 2 scoring strategies. The first strategy comprises the end-to-end training of models pretrained on ImageNet.[7,17] A 3-stage model was developed to output the results of intersected cancer and nuclear segmentation masks for counting viable cells. Further superimposing of a positive pixel mask identified PD-L1-positive cells for calculations of TPS.[7] Considering the difference in scale between cancer nests and nuclei, which are both segmented by the algorithm, a dual-scale categorization-based deep learning method was proposed, which employed 2 separate VGG16 neural networks for high and low magnification. This method showed a concordance of 88% with pathologists, which was higher than the 83% concordance of a 1-scale categorization-based method.[18] The second strategy for PD-L1 TPS generation consists of transfer learning models together with supervised machine learning steps using handcrafted features. An example of using this strategy is PD-L1 scoring with the open-source QuPath application, which provided good concordance with pathologists.[19] Built-in cell detection and classification functions in QuPath were used to score PD-L1 levels in urothelial carcinoma and resulted in a correlation with pathologists of $r = 0.834$ ($P < .001$).[19] The latest study using multicentric and multi-PD-L1 assay data showed moderately low agreement at cell level PD-L1 expression, compared with 6 pathologists. However, the agreement on TPS quantification using ICC is at 0.796 (95% CI: 0.694-0.898).[20] Our pipeline in Visiopharm also employs a framework of transfer learning combined with steps that rely on handcrafted features. It utilizes DeepLab V3 as the basic model for cancer segmentation. A DeepLab V3 framework was also used by others in the regional segmentation model, R-Net, and further combined with C-Net to develop an automated TPS algorithm. Surprisingly, the agreement between digital and mTPS scores was higher at low compared with high cutoff TPS values,[8] possibly related to the multiscale nature of this model segmenting both cancer nests and nuclei. Although the studies demonstrate the feasibility of using both de novo training and transfer learning for digital TPS generation, they did not evaluate the role of cancer segmentation in the overall accuracy of TPS.

Our study focuses on optimizing a workflow that is transparent and reduces the annotation time of pathologists. As a starting point, we use coregistered 40$\times$ tiles from H&E, IHC, and NC images, which are obtained from 3 parallel slides of the same tissue block. Pathologists selected 3 small cancer tiles (1028 $\times$ 1028 pixels) of the same size in the WSI of the PD-L1 digital slide. Pathologists also provided manual outlines of cancer regions in PD-L1 and NC tiles, which is the most time-consuming task in the project. H&E slides and tiles are available to help with difficult cancer diagnoses. The rest of the workflow does not require the domain expertise of a pathologist and, in our case, utilizes modules in Visiopharm software. However, the same modules also exist in other software packages, such as in the open-source QuPath package, making the workflow broadly applicable.

CLIA/CAP issued a recommendation in 2021 for digital primary diagnosis.[21] Guidelines for IHC are included in conjunction with the primary diagnosis, which pertains primarily to increasing the diagnostic accuracy of IHC. CLIA/CAP mandates a sample set of at least 60 cases for 1 tissue preparation, for example, H&E-stained sections of fixed tissue or frozen sections that are representative of routine practice cases. They also recommend the inclusion of another 20 cases such as IHC or other special stains if these stains are needed for diagnosis and not included in the 60 cases. Because reporting of IHC results does not fall under the purview of primary diagnosis, case numbers for algorithmic IHC LDT validation are not specified. Extending the recommendations for primary diagnosis, a set of 60 to 80 cases would be considered adequate for algorithmic IHC LDT validation. The cases would then be divided into training and test sets. Our data suggest that for validation of a PD-L1 IHC quantification algorithm on cases that were stained at the same site but originated at multiple different hospitals, a set of ~75 cases may be sufficient for algorithmic LDT validation. Of those, 15

cases are used for training and 60 cases for testing of the TPS algorithm. However, we would like to caution that the exact number of cases may depend on the cancer type and the IHC stain. In addition, the accuracy of our pipeline does not reach the 95% concordance between manual and algorithmic results that are required according to CLIA/CAP.

The project has several limitations. (1) We enrolled cases based on the size of tissue that was used for PD-L1 IHC. This excluded 60% of lung cancer cases that were diagnosed based on endobronchial biopsy or fine needle aspiration. Thus, there exists an urgent need to develop algorithms for other types of tissue preparation. (2) We are using a pretrained CNN model as a starting point for training of a cancer segmentation model. Recently, other pretrained models that utilize Transformer network architectures and are trained on millions of pathology images have become available.[22-24] Applying these models to the cancer segmentation task will likely improve the algorithmic performance. (3) In our proposed framework, parameters were adjusted for nuclear segmentation using the rule-based model in Visiopharm software. A later version of the Visiopharm software includes deep learning nuclear segmentation models that provide more consistent and accurate nuclear outlines. (4) A major limitation of the current proposed model is its confusion of tumor and immune cells that are adjacent to the tumor,[25] in particular, under conditions where both cell types are PD-L1-positive. Models therefore rely on nuclear features for cancer cell segmentation, which are not affected by PD-L1 staining.[26] The training of cell type-specific models can be improved by antibody staining with different chromogens for immune and cancer cells preferably in the same tissue section to avoid coregistration errors.[27,28]

In conclusion, we propose a systematic workflow to train and test multistep models for analysis of tissues stained by IHC. IHC tumor markers that are expressed in both cancer and cells within the tumor microenvironment require the separation of tumor regions from intervening benign areas. The accuracy of tumor segmentation affects the overall accuracy of computer-assisted IHC scoring solutions. The separation of an IHC scoring pipeline into multiple steps allows us to determine which steps require optimization and increase the transparency of the algorithmic scoring system. In addition, a systematic process to evaluate the accuracy of the IHC scoring pipeline increases the trust of end-user pathologists and facilitates the deployment of software for clinical practice.

*Acknowledgments*

The study was conducted under the institutional IRB protocol #00091019. We thank the staff in the ARUP ICL laboratory for help with slide scanning. We thank the Department of Pathology, Biorepository and Molecular Pathology (BMP) Core at the Huntsman Cancer Institute (salary support for WZ) and ARUP for supporting the work of this study.

*Author Contributions*

B.K. and G.D. designed the concept, A.J., W.Z., and L.J.P. performed and collected the data, J.Y. and B.B. provided the statistical analysis, B.S.K. and W.Z. wrote the manuscript, B.S.K. and J.T. revised the manuscript, all authors read and approved the final version of the manuscript.

*Data Availability*

The data that support the findings of this study are available from the corresponding author (B.S.K. and W.Z.), upon reasonable request.

*Funding*

No outside funding was provided.

*Declaration of Competing Interest*

None reported.

## Supplementary Material

The online version contains supplementary material available at https://doi.org/10.1016/j.labinv.2024.102070

## References

1. Kim SW, Roh J, Park CS. Immunohistochemistry for pathologists: protocols, pitfalls, and tips. *J Pathol Transl Med*. 2016;50(6):411–418.
2. Akhtar M, Rashid S, Al-Bozom IA. PD–L1 immunostaining: what pathologists need to know. *Diagn Pathol*. 2021;16(1):94.
3. Han Y, Liu D, Li L. PD-1/PD-L1 pathway: current researches in cancer. *Am J Cancer Res*. 2020;10(3):727–742.
4. Cha JH, Chan LC, Li CW, Hsu JL, Hung MC. Mechanisms controlling PD-L1 expression in cancer. *Mol Cell*. 2019;76(3):359–370.
5. Troncone G, Gridelli C. The reproducibility of PD-L1 scoring in lung cancer: can the pathologists do better? *Transl Lung Cancer Res*. 2017;6(Suppl 1):S74–S77.
6. Baxi V, Lee G, Duan C, et al. Association of artificial intelligence-powered and manual quantification of programmed death-ligand 1 (PD-L1) expression with outcomes in patients treated with nivolumab ± ipilimumab. *Mod Pathol*. 2022;35(11):1529–1539.
7. Liu J, Zheng Q, Mu X, et al. Automated tumor proportion score analysis for PD-L1 (22C3) expression in lung squamous cell carcinoma. *Sci Rep*. 2021;11(1):15907.
8. Pan B, Kang Y, Jin Y, et al. Automated tumor proportion scoring for PD-L1 expression based on multistage ensemble strategy in non-small cell lung cancer. *J Transl Med*. 2021;19(1):249.
9. Puladi B, Ooms M, Kintsler S, et al. Automated PD-L1 scoring using artificial intelligence in head and neck squamous cell carcinoma. *Cancers (Basel)*. 2021;13(17):4409.
10. Song L, Zeng L, Yan H, et al. Validation of E1L3N antibody for PD-L1 detection and prediction of pembrolizumab response in non-small-cell lung cancer. *Commun Med (Lond)*. 2022;2(1):137.
11. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163.
12. Kapil A, Meier A, Zuraw A, et al. Deep semi supervised generative learning for automated tumor proportion scoring on NSCLC tissue needle biopsies. *Sci Rep*. 2018;8(1), 17343.
13. Lin D, Wu J, Sun W, et al. MA15.02 deep learning approach for automated tumor cells detection and estimation of PD-L1 22C3 assay expression in lung adenocarcinoma. *J Thorac Oncol*. 2019;14(10):S309.
14. Hondelink LM, Huyuk M, Postmus PE, et al. Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumour proportion score assessment in non-small cell lung cancer. *Histopathology*. 2022;80(4):635–647.
15. Cheng G, Zhang F, Xing Y, et al. Artificial intelligence-assisted score analysis for predicting the expression of the immunotherapy biomarker PD-L1 in lung cancer. *Front Immunol*. 2022;13:893198.
16. Prizant H, Shamshoian J, Abel J, et al. Abstract 5358: digital SP263 PD-L1 tumor cell scoring in non-small cell lung cancer achieves comparable outcome prediction to manual pathology scoring. *Cancer Res*. 2023;83(7_Supplement), 5358-5358.
17. Wu J, Liu C, Liu X, et al. Artificial intelligence-assisted system for precision diagnosis of PD-L1 expression in non-small cell lung cancer. *Mod Pathol*. 2022;35(3):403–411.
18. Wang X, Chen P, Ding G, et al. Dual-scale categorization based deep learning to evaluate programmed cell death ligand 1 expression in non-small cell lung cancer. *Medicine (Baltimore)*. 2021;100(20):e25994.

19. Rodrigues A, Nogueira C, Marinho LC, et al. Computer-assisted tumor grading, validation of PD-L1 scoring, and quantification of CD8-positive immune cell density in urothelial carcinoma, a visual guide for pathologists using QuPath. *Surg Exp Pathol*. 2022;5(1):12.
20. van Eekelen L, Spronck J, Looijen-Salamon M, et al. Comparing deep learning and pathologist quantification of cell-level PD-L1 expression in non-small cell lung cancer whole-slide images. *Sci Rep*. 2024;14(1):7136.
21. Evans AJ, Brown RW, Bui MM, et al. Validating whole slide imaging systems for diagnostic purposes in pathology. *Arch Pathol Lab Med*. 2022;146(4):440–450.
22. He K, Gan C, Li Z, et al. Transformers in medical image analysis. *Intell Med*. 2023;3(1):59–78.
23. Barzekar H, Patel Y, Tong L, Yu Z. MultiNet with transformers: a model for cancer diagnosis using images. *arXiv preprint arXiv*:230109007. 2023.
24. Chen RJ, Chen C, Li Y, et al. *Scaling vision transformers to gigapixel images via hierarchical self-supervised learning*. 2022:16144–16155.
25. Gordon SR, Maute RL, Dulken BW, et al. PD-1 expression by tumour-associated macrophages inhibits phagocytosis and tumour immunity. *Nature*. 2017;545(7655):495–499.
26. Lal S, Das D, Alabhya K, Kanfade A, Kumar A, Kini J. NucleiSegNet: Robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images. *Comput Biol Med*. 2021;128:104075.
27. Komura D, Onoyama T, Shinbo K, et al. Restaining-based annotation for cancer histology segmentation to overcome annotation-related limitations among pathologists. *Patterns (N Y)*. 2023;4(2):100688.
28. Ma Z, Shiao SL, Yoshida EJ, et al. Data integration from pathology slides for quantitative imaging of multiple cell types within the tumor immune cell infiltrate. *Diagn Pathol*. 2017;12(1):69.