

Research Article

A Deep Learning—Based Assay for Programmed Death Ligand 1 Immunohistochemistry Scoring in Non—Small Cell Lung Carcinoma: Does it Help Pathologists Score?

Hiroaki Ito^a, Akihiko Yoshizawa^{a,b,*}, Kazuhiro Terada^a, Akiyoshi Nakakura^c, Mariyo Rokutan-Kurata^a, Tatsuhiko Sugimoto^d, Kazuya Nishimura^d, Naoki Nakajima^{a,e}, Shinji Sumiyoshi^{a,f}, Masatsugu Hamaji^g, Toshi Menju^g, Hiroshi Date^g, Satoshi Morita^c, Ryoma Bise^d, Hironori Haga^a

^a Department of Diagnostic Pathology, Kyoto University Hospital, Kyoto, Japan; ^b Department of Diagnostic Pathology, Nara Medical University, Nara, Japan; ^c Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan; ^d Department of Advanced Information Technology, Kyushu University, Fukuoka, Japan; ^e Department of Diagnostic Pathology, Toyooka Hospital, Hyogo, Japan; ^f Department of Diagnostic Pathology, Tenri Hospital, Nara, Japan; ^g Department of Thoracic Surgery, Kyoto University Hospital, Kyoto, Japan

ARTICLE INFO

Article history:

Received 10 August 2023

Revised 8 February 2024

Accepted 1 April 2024

Available online 6 April 2024

Keywords:

AI-assisted system

deep learning-based assay

non-small cell lung carcinoma

PD-L1

immunohistochemistry

ABSTRACT

Several studies have developed various artificial intelligence (AI) models for immunohistochemical analysis of programmed death ligand 1 (PD-L1) in patients with non-small cell lung carcinoma; however, none have focused on specific ways by which AI-assisted systems could help pathologists determine the tumor proportion score (TPS). In this study, we developed an AI model to calculate the TPS of the PD-L1 22C3 assay and evaluated whether and how this AI-assisted system could help pathologists determine the TPS and analyze how AI-assisted systems could affect pathologists' assessment accuracy. We assessed the 4 methods of the AI-assisted system: (1 and 2) pathologists first assessed and then referred to automated AI scoring results (1, positive tumor cell percentage; 2, positive tumor cell percentage and visualized overlay image) for final confirmation, and (3 and 4) pathologists referred to the automated AI scoring results (3, positive tumor cell percentage; 4, positive tumor cell percentage and visualized overlay image) while determining TPS. Mixed-model analysis was used to calculate the odds ratios (ORs) with 95% CI for AI-assisted TPS methods 1 to 4 compared with pathologists' scoring. For all 584 samples of the tissue microarray, the OR for AI-assisted TPS methods 1 to 4 was 0.94 to 1.07 and not statistically significant. Of them, we found 332 discordant cases, on which the pathologists' judgments were inconsistent; the ORs for AI-assisted TPS methods 1, 2, 3, and 4 were 1.28 (1.06-1.54; $P = .012$), 1.29 (1.06-1.55; $P = .010$), 1.28 (1.06-1.54; $P = .012$), and 1.29 (1.06-1.55; $P = .010$), respectively, which were statistically significant. For discordant cases, the OR for each AI-assisted TPS method compared with the others was 0.99 to 1.01 and not statistically significant. This study emphasized the usefulness of the AI-assisted system for cases in which pathologists had difficulty determining the PD-L1 TPS.

© 2024 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights reserved.

Introduction

Lung cancer is one of the most common malignant tumors worldwide.¹ Immune checkpoint inhibitors (ICIs) targeting

* Corresponding author.

E-mail address: akyoshi@kuhp.kyoto-u.ac.jp (A. Yoshizawa).



programmed death 1 or programmed death ligand 1 (PD-L1) result in better survival of patients with various solid tumors, including lung cancer, compared with standard treatment options,^{2,3} offering a personalized and less invasive alternative therapy.⁴ Pembrolizumab, an ICI, has been indicated for the treatment of non-small cell lung carcinoma (NSCLC)^{5,6} and approved worldwide as a combination therapy and first-line treatment option for patients with advanced NSCLC.^{7,8}

Following clinical trial reports, the US Food and Drug Administration has approved the PD-L1 immunohistochemistry (IHC) assay (22C3; DAKO, Agilent) as a companion diagnostic tool for PD-L1 immunotherapy with pembrolizumab.⁹ The main manufacturer of 22C3 assays proposed a guideline for 22C3 staining assessment by estimating the tumor proportion score (TPS).¹⁰ However, this scoring guideline is based on descriptions and examples and is therefore subjective for calculating the TPS. Scoring variability among pathologists could be problematic when assessing PD-L1 expression.¹¹⁻¹⁴ Consequently, an objective, repeatable, and accurate PD-L1 evaluation method is warranted.

Digital image analysis is expected to overcome these limitations. Artificial intelligence (AI) techniques, including deep learning (DL) algorithms, can help pathologists improve diagnostic accuracy, concordance, and efficiency.¹⁵⁻¹⁹ Specifically, several DL models for PD-L1 IHC analysis have been developed for NSCLC.²⁰⁻²⁶ Previous studies have focused on the correlation coefficient between pathologists' and automated scores and concluded that the automated PD-L1 scoring system showed accuracy similar to that of pathologists.^{20-23,25-27} Recently, Wu et al.²⁴ reported that an AI-assisted diagnostic system for evaluating PD-L1 expression could improve the diagnostic accuracy and efficiency of pathologists. However, none of these studies have focused on the specific ways by which AI-assisted systems could help pathologists determine the TPS. In the field of radiology, several authors have conducted comparative studies evaluating 2 types of computer-aided detection (CAD) systems on computed tomography (CT). Considering the concurrent-reader mode, the radiologist reads the CT images while viewing the CAD results. Regarding the second-reader mode, the radiologist first reads CT images alone and then inspects the CAD.²⁸⁻³⁰ To the best of our knowledge, no comparative studies have been conducted on how an AI system could assist pathologists.

Based on this background, we planned an observational study to generate a DL-based pipeline to calculate the TPS of the 22C3 assay and evaluate it in comparison with pathologists' results using whole-slide images (WSIs) in patients with NSCLC. Furthermore, we analyzed how the AI-assisted system could affect the diagnostic accuracy of the pathologists' assessment.

Materials and Methods

Patient Selection

The patient selection criteria were as follows: (1) a histologic diagnosis of NSCLC, (2) resection at Kyoto University Hospital from 2001 to 2015, and (3) no systemic chemoradiotherapy before surgery. Data on clinicopathologic patient characteristics, including age, sex, smoking history, and pathologic T level,³¹ were collected. All cases were reviewed and classified into subtypes by pathologists (N.N. and A.Y.) according to the 2021 World Health Organization criteria.³² The need for consent was waived for this retrospective study, which assessed pathologic specimens with limited clinical

information. This retrospective study was approved by the Institutional Ethics Committee (reference number: R1876-1).

Construction of Tissue Microarrays

All resected NSCLC specimens were fixed in formalin solution, sectioned, and stained with hematoxylin and eosin (H&E) conventionally. Representative tumor regions without showing necrosis were selected and marked on H&E-stained slides for each case. Briefly, a representative core (2 mm in diameter) with sufficient tissue quality and amount was stamped out of the donor block and transferred to a recipient paraffin block. Serial 4- μ m sections were used for H&E staining and subsequent IHC. In total, the following 692 samples were included: 665 NSCLC samples and 27 normal lung tissue samples (Fig. 1).

Immunohistochemistry and Digitization

Serial sections of the tissue microarrays (TMAs) constructed from the samples were IHC stained using the Dako pharmDx kits on the Dako Link 48 platform (DAKO, Agilent) for clones 22C3, according to the manufacturer's recommendations and counterstained with hematoxylin. All sections were scanned at $\times 40$ magnification (equivalent to $\times 400$ in light microscopy) using NanoZoomer 2.0-HT (Hamamatsu Photonics).

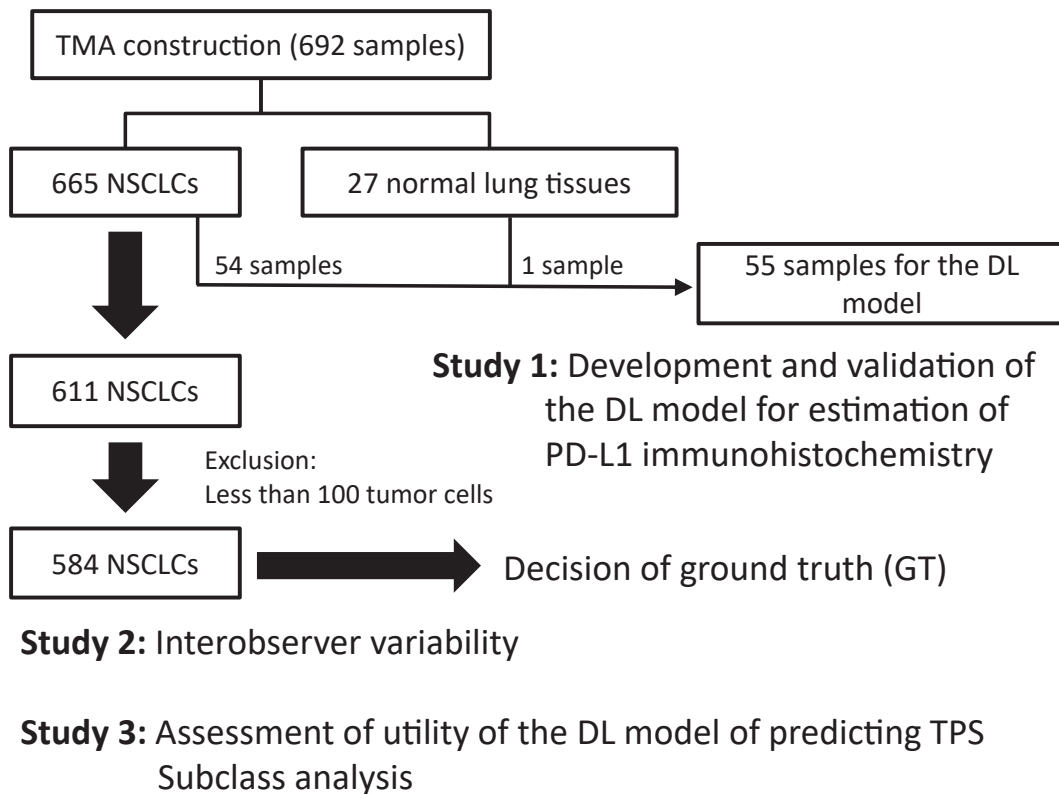
Study 1: Development and Validation of a Deep Learning Model to Estimate PD-L1 Immunohistochemistry

Collection of Training and Validation Data Sets and Annotation

Fifty-four carcinoma samples and 1 normal lung tissue sample from 692 samples were randomly selected (Fig. 1). A pathologist (H.I.; Computation Pathology Group, Diagnostic Image Analysis Group, Radboud University Medical Center) manually annotated them on a WSI using the annotation software ASAP, version 1.9.0, according to the manual on "PD-L1 immunohistochemistry testing in lung cancer" reported by the International Association for the Study of Lung Cancer.³³ Each annotated region on the WSI included only a single class of cells: class 0, area of tumor cells without PD-L1 expression; class 1, area of tumor cells with PD-L1 expression; and class 2, area of nontumor cells (Supplementary Fig. S1).

Development of a Deep Learning Model Pipeline

A DL model was developed with 2 parts: (1) a convolutional network for nuclei recognition and (2) a regional segmentation network (Fig. 2). Regarding nuclei recognition, all nuclei in the core were extracted by learning using U-net and classified into 2 categories: tumor cells and nontumor cells.³⁴ We used DeepLabv3³⁵ for the regional segmentation network pretrained by ImageNet. The network could estimate 2 classes (negative or positive) for each pixel. Multilabel soft margin loss³⁶ was used for the loss function. The minibatch size was 8; the number of epochs was 100. The Adam algorithm was used as the optimizer with a 1e-3 learning rate. In inference, given a patch image in a WSI image as the input, the nuclei recognition network detected the positions of tumor cells; the regional segmentation network segmented the region as PD-L1 positive or negative. Subsequently, the system detected the positions of the 3 types of cells, nontumor, PD-L1-positive tumor, and PD-L1-negative tumor cells, by combining these 2 results. Using the detected results, the PD-L1-positive rate was calculated.

**Figure 1.**

A flowchart of the study. This study includes 3 parts: study 1, development and validation of the deep learning model for estimation of PD-L1 immunohistochemistry; study 2, evaluation of interobserver variability among the pathologists; and study 3, assessment of the utility of the deep learning model for predicting the tumor proportion score and subclass analysis. DL, deep learning; NSCLC, non-small cell lung carcinoma; PD-L1, programmed death ligand 1; TMA, tissue microarray; TPS, tumor proportion score.

Experiment Settings

To demonstrate the cell classification ability, we conducted 3-fold cross-validation, which was as follows: 55 cores (350 patch images) captured using a WSI scanner were divided into 3 sets. The cores were partially annotated as ground truth (GT) regions; only the annotated regions were used for cross-validation. We divided the cores such that the annotated regions in each class were balanced. Regarding the test, a detected cell was assigned to 1 of the 3 classes (nontumor, positive tumor, or negative tumor) using our AI model. For each class (called the target class), such as a positive tumor, if a cell was correctly estimated as a target class, it was counted as true positive. If a cell belonging to another class was mistakenly estimated to belong to the target class, it was counted as false positive. If a cell belonging to the target class was mistakenly estimated to belong to another class, it was counted as false negative. We evaluated the model performance using 3 metrics: the mean of the per-class precision (mPrecision), the mean of the per-class recall (mRecall), and the F1 score (F1). These are defined as follows:

$$\text{mPrecision} = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{c=1}^3 TP_i^c}{\sum_{c=1}^3 TP_i^c + \sum_{c=1}^3 FP_i^c}$$

$$\text{mRecall} = \frac{1}{N} \sum_{i=1}^N \frac{\sum_{c=1}^3 TP_i^c}{\sum_{c=1}^3 TP_i^c + \sum_{c=1}^3 FN_i^c}$$

$$F1 = \frac{2 \times \text{mPrecision} \times \text{mRecall}}{\text{mPrecision} + \text{mRecall}},$$

in which N is the number of annotated cells.

Study 2: Interobserver Variability of Tumor Proportion Scoring

Final Objective Cores

We excluded 55 of the 692 cores because they were used to develop the DL model. In addition, we excluded 27 cores with <100 tumor cells (Fig. 1).

Assessment of PD-L1 Expression by Pathologists and Tumor Proportion Score (Ground Truth) Determination

Three board-certified pathologists (H.I., K.T., and A.Y.) participated in the assessment of PD-L1 expression; H.I. and K.T. had 8 years of experience, and A.Y. had >23 years of experience. The assessment was based on a previously described manual.³³ Scores were recorded as the percentage of PD-L1-positive tumor cells over the total number of tumor cells and the TPS. Staining status was classified into the following categories: <1%, ≥1% and <50%, and ≥50%. Each pathologist determined the TPS for 584 cases thrice at appropriate intervals (>2 weeks) without knowledge of their own results and those of other pathologists. Subsequently, the GT was determined based on a consensus between the 3 pathologists as a 3-category scoring for 584 cases. Furthermore, interobserver variability of the 3 pathologists' scoring was calculated for 3-tiered assessment (<1%, ≥1% and <50%, and ≥50%), 1%

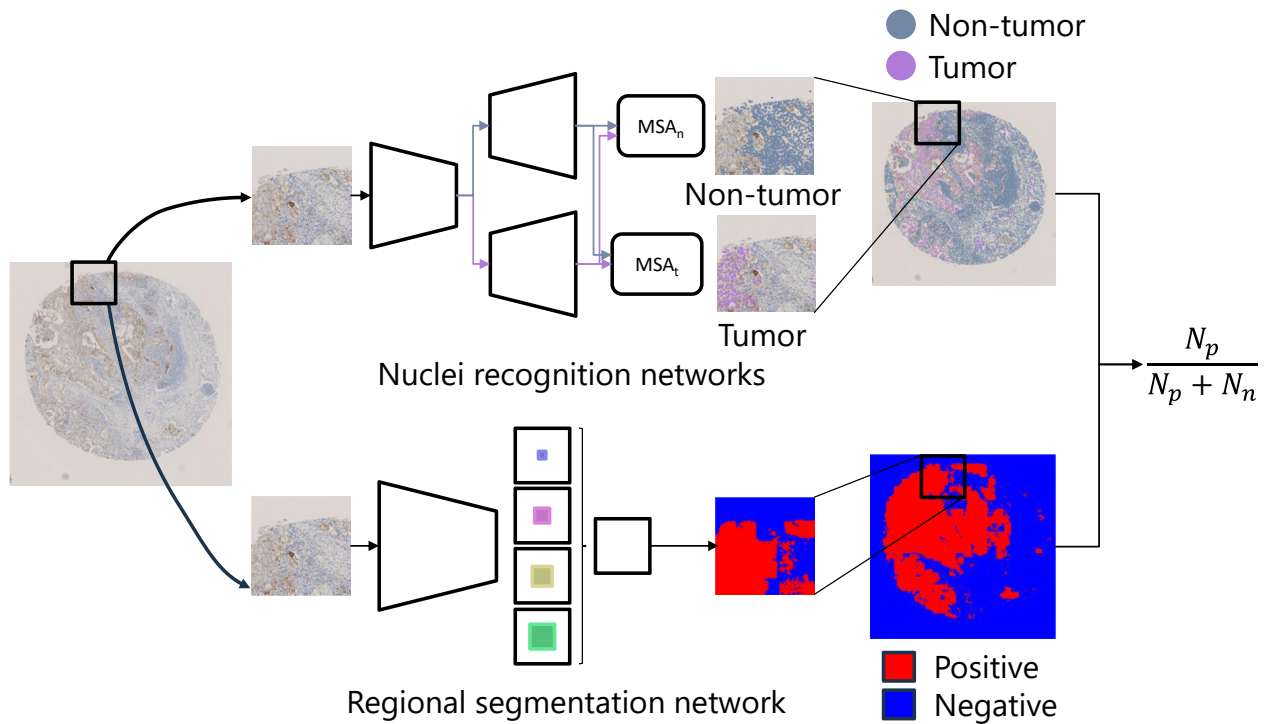


Figure 2.

A pipeline of the deep learning model. The deep learning model estimates tumor or nontumor cells (nuclei recognition using convolutional networks) and PD-L1 positivity or negativity (regional segmentation network). By combining these 2 results, the system detects PD-L1–positive and PD-L1–negative tumor cells and calculates the tumor proportion score.

MSA_n, modified self-attention for nontumor cells; MSA_t, modified self-attention for tumor cells; N_n, number of negative tumor cells; N_p, number of positive tumor cells; PD-L1, programmed death ligand 1.

cutoff 2-tiered assessment (<1% and ≥1%), and 50% cutoff 2-tiered assessment (<50% and ≥50%) as the average of the 3 times scoring.

Study 3: Assessment of Utility of the Deep Learning Model for Predicting the Tumor Proportion Score

Tumor Proportion Scoring Using the Deep Learning Model

First, the TPS was estimated using the DL model developed for 584 cases. The results of the DL model are shown as the percentage of PD-L1–positive tumor cells/tumor cells (continuous variable). These results were visualized using overlay images displaying PD-L1–positive tumor, PD-L1–negative tumor, and nontumor cells (Fig. 3). Agreement rates with the GT were calculated for each pathologist's scoring without AI (Path-TPS) and for automated scoring using the DL model (AI-TPS).

Tumor Proportion Scoring in an Artificial Intelligence–Assisted System

Figure 4 shows TPS determination. The pathologists scored the samples according to the AI-TPS without knowledge of their own results and those of other pathologists after an appropriate washout period (>2 weeks). TPS was evaluated using 4 methods: (1) the pathologists first determined the TPS and then referred to the AI-TPS results (percentage) for final confirmation; (2) the pathologists first determined the TPS and then referred to the AI-TPS results (percentage and overlay image) for final confirmation; (3) the pathologists referred to the AI-TPS results (percentage) while determining the TPS; and (4) the pathologists referred to the AI-TPS results (percentage and overlay image) while

determining the TPS. Methods 1 and 2 correspond to the AI-assisted system in the second-reader mode and methods 3 and 4 correspond to the concurrent-reader mode in Figure 4. The scoring results for each method were compared with those of the GT, and the agreement rates were calculated. Using a mixed-model analysis, we calculated the odds ratios (ORs) with 95% confidence intervals (CIs) for AI-assisted TPS methods 1, 2, 3, and 4 and compared them with those for Path-TPS.

Subclass Analysis of Tumor Proportion Scoring

Regarding subclass analysis, 584 cases were divided into the following subclasses: (1) adenocarcinomas, (2) squamous cell carcinomas, and (3) discordant cases. In this study, 3 pathologists separately determined the TPS 3 times at appropriate intervals without knowledge of their own results, those of other pathologists, or those of the AI-TPS. The total count of judgments was 9. Concordant cases were defined as cases in which the 9 judgments were in perfect agreement. Other cases were defined as discordant cases.

Study 4: Assessment of the Artificial Intelligence–Assisted System Using Surgical Specimens

We performed further assessment of the AI-assisted system for evaluating surgical specimens to investigate whether it would apply not only to TMA slides but also to actual clinical practice. In the patient cohort of resection cases at Kyoto University Hospital from 2001 to 2016, we used 47 available slides stained with PD-L1 (22C3) as we previously described, which are as follows: 37

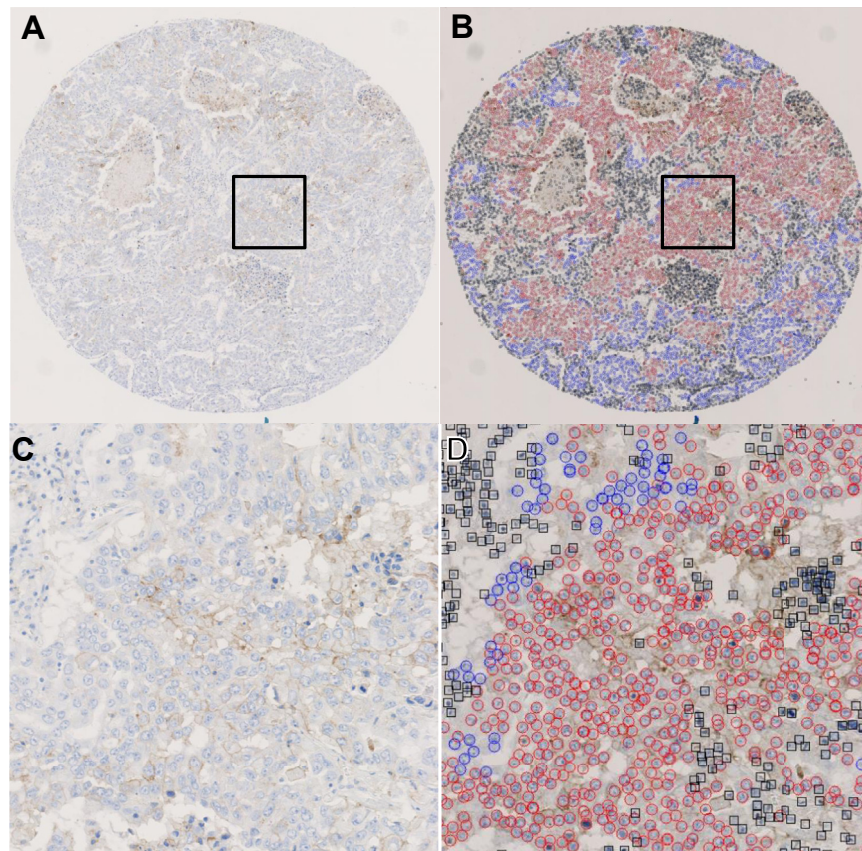


Figure 3.

Representative images of the visualized overlay image. (A) Programmed death ligand 1 (PD-L1) immunostaining of representative sample. (B) The visualized overlay image of the same sample. Red circles indicate PD-L1-positive tumor cells; blue circles indicate PD-L1-negative tumor cells; black squares indicate nontumor cells. (C) The squared area in panel A. (D) The squared area in panel B.

adenocarcinoma, 8 squamous cell carcinoma, 1 adenosquamous carcinoma, and 1 pleomorphic carcinoma cases. All slides were scanned at $\times 40$ magnification (equivalent to $\times 400$ in light microscopy) and annotated with the tumor area. Using the DL model, the TPS of the annotated tumor area was calculated as a continuous variable for each case. The 3 pathologists categorized the TPS without AI (Path-TPS) into 3 categories: $<1\%$, $\geq 1\%$ and $<50\%$, and $\geq 50\%$. After an appropriate washout period (>2 weeks), they determined the GT of the 47 cases based on a consensus among them in the 3 categories. Then, after an appropriate washout period (>2 weeks), they determined the TPS using the AI-assisted system (AI-assisted TPS) in the 3 categories. A representative method (the third method, concurrent-reading mode without referring to the overlay visualized image) was chosen. Agreement rates with the GT were calculated for AI-TPS, Path-TPS, and AI-assisted TPS. We performed subclass analysis of the discordant cases in which the judgments of TPS by each pathologist, without reference to the AI-TPS, did not perfectly agree, as described in the TMA study.

Statistical Analyses

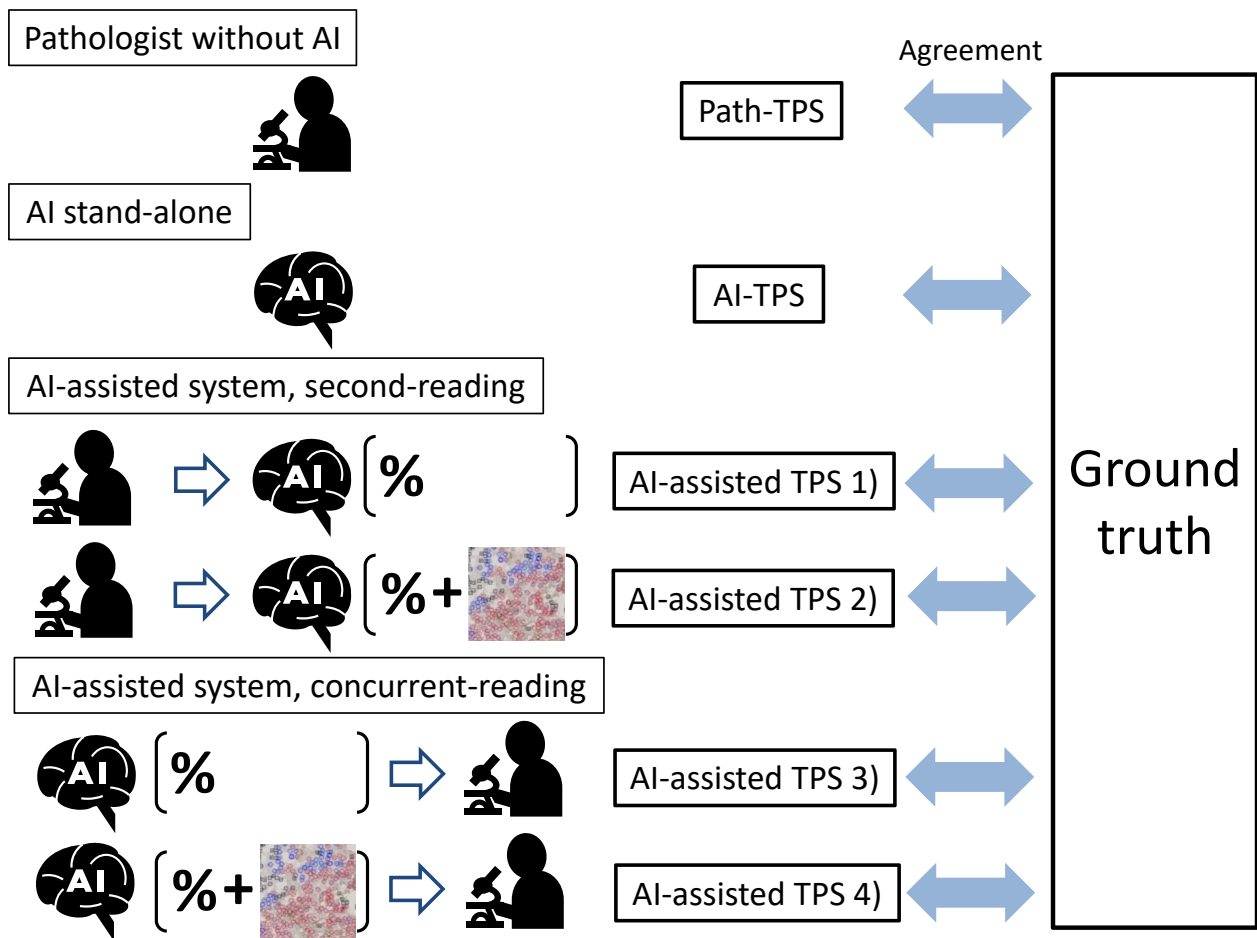
The Fleiss κ coefficient for the overall agreement between all observer scores without AI was calculated. The level of agreement based on the κ statistics was defined as follows: $\kappa < 0.20$ as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial, and >0.81 as almost perfect agreement.³⁷ Statistical

analyses were performed using the R software version 4.1.0 (R Project for Statistical Computing). The κ score was calculated using the R package “irr.” A mixed-effect logistic model analysis was performed to analyze the proportional concordance with the GT. The mixed model used the patterns of AI as a fixed effect and cases and pathologists as random effects. Using a mixed-model analysis, we calculated the ORs with 95% CIs for AI-assisted TPS methods 1, 2, 3, and 4 in comparison with the Path-TPS. We calculated each value for the total cases and the following subclass groups: adenocarcinomas, squamous cell carcinomas, and discordant cases. Furthermore, for discordant cases, we calculated the OR with 95% CI for each AI-assisted TPS in comparison with the others in all pairs of the 4 methods. The tests were 2-sided, and P values $< .05$ were considered statistically significant.

Results

Case Demographics

The baseline characteristics are shown in Table 1. Of the 584 cores, 278 adenocarcinoma, 292 squamous cell carcinoma, 1 adenosquamous carcinoma, and 13 non-small cell carcinoma cores were used for prediction using the DL model. Of the 584 cores, 362, 122, and 100 cases were classified as TPS $<1\%$ cases; $\geq 1\%$ and $<50\%$ cases; and $\geq 50\%$ cases, respectively. Of the 278 cases of adenocarcinoma, 197, 54, and 27 cases were classified as $<1\%$ cases; $\geq 1\%$ and $<50\%$ cases; and $\geq 50\%$ cases, respectively. Of

**Figure 4.**

A schematic of the study. Path-TPS means scoring by a pathologist without AI. AI-TPS refers to scoring using AI alone. The AI-assisted TPS methods include 1, 2, 3, and 4. Methods 1 and 2 correspond to the second-reading mode, and methods 3 and 4 correspond to the concurrent-reading mode. In methods 1 and 3, the pathologist refers to the percentage predicted by the AI and in methods 2 and 4, the percentage and visualized overlay image. Each TPS is compared with the ground truth to assess the agreement. AI, artificial intelligence; TPS, tumor proportion score.

the 292 cases of squamous cell carcinoma, 158, 68, and 66 cases were classified as <1% cases; $\geq 1\%$ and <50% cases; and $\geq 50\%$ cases, respectively.

Study 1: 3-Fold Cross-Validation of the Deep Learning Model

We performed 3-fold cross-validation by dividing the annotated cores such that each area of the 3 classes was equal. Each pixel in the 3 folds is listed in [Supplementary Table S1](#). The results of the validation (mPrecision, mRecall, and F1) are summarized in [Supplementary Table S2](#). The average values of mPrecision, mRecall, and F1 were 0.686, 0.857, and 0.762, respectively.

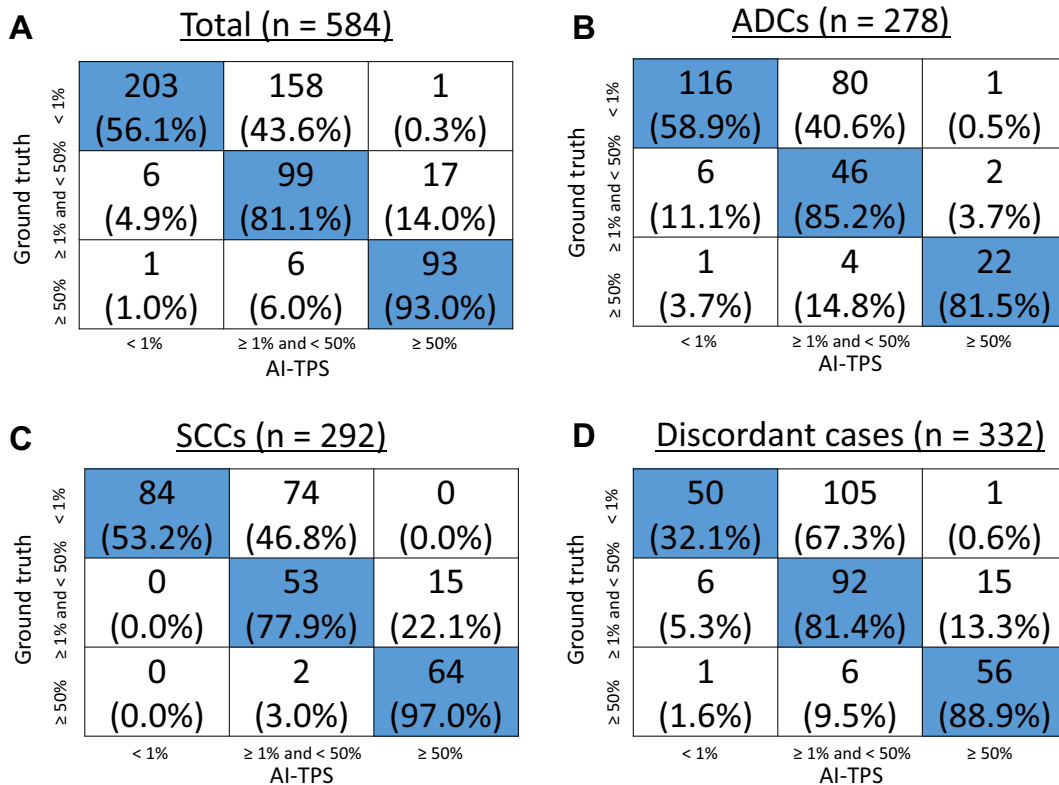
Study 2: Interobserver Variability of Tumor Proportion Scoring by Pathologists (Fleiss' K Coefficient)

The interobserver variability among the 3 pathologists was calculated for the Path-TPS. The κ coefficients for 3-tiered, 1% cutoff 2-tiered, and 50% cutoff 2-tiered assessments were 0.581, 0.619, and 0.794, respectively. These results are presented in [Supplementary Table S3](#), in comparison with previous reports.

Table 1
Case demographics

Characteristic	Total	ADC	SCC	Others
Total	584	278	292	14
Age (y)				
Average	68.3	65.7	70.9	64.8
Range	23–88	23–88	37–85	46–79
Sex				
Male	412	146	254	12
Female	172	132	38	2
Stage (AJCC 8th)				
Tis	3	2	1	0
T1	267	165	101	1
T2	209	91	111	7
T3	63	11	47	5
T4	42	9	32	1
Smoking history				
Current/former smoker	442	150	280	12
Never smoker	142	128	12	2
TPS				
1%	362	197	158	7
$\geq 1\%$ and < 50%	122	54	68	0
$\geq 50\%$	100	27	66	7

ADC, adenocarcinoma; AJCC, American Joint Committee on Cancer; SCC, squamous cell carcinoma; TPS, tumor proportion score.

**Figure 5.**

Confusion matrices of AI-TPS. Confusion matrices for (A) 584 cases; (B) adenocarcinomas (278 cases); (C) squamous cell carcinomas (292 cases); and (D) discordant cases (332 cases). ADCs, adenocarcinomas; AI, artificial intelligence; SCCs, squamous cell carcinomas; TPS, tumor proportion score.

Study 3: Confusion Matrices of Tumor Proportion Scoring Using the Deep Learning Model

The confusion matrix for the AI-TPS is shown in [Figure 5A](#). Each value in the matrix is presented as a percentage of the total number of true labels. Of the 584 cases, 67.6% (395/584) were matched for GT. The confusion matrices for each subclass are shown in [Figure 5B-D](#). Of the 278 cases of adenocarcinoma, 292 cases of squamous cell carcinoma, and 332 discordant cases, 66.2% (184/278), 68.8% (201/292), and 59.6% (198/332) were matched, respectively.

Agreement (Overall Accuracy) Between the Ground Truth and the Artificial Intelligence Tumor Proportion Score/Pathologists' Tumor Proportion Score/Artificial Intelligence–Assisted Tumor Proportion Score

The detailed agreement (overall accuracy) of the AI-TPS, Path-TPS, and AI-assisted TPS listed are shown in [Supplementary Table S4](#). In all the 584 cases, the agreements of the Path-TPS of all pathologists, calculated as the average of 3 times scoring (75.8%–83.6%), were higher than those of AI-TPS (67.6%). The agreements of AI-assisted TPS of all 4 methods (73.3%–87.2%) were higher than those of AI-TPS (67.6%). A similar trend was observed in the subclasses of adenocarcinomas and squamous cell carcinomas. Regarding the discordant subclass, the agreement of the Path-TPS of the 2 pathologists (71.1%–64.5%) was higher than that of the AI-TPS (59.6%), whereas the agreement of the Path-TPS of the others (57.4%) was lower. In contrast, the agreements of AI-assisted TPS of all 4 methods of all pathologists (59.9%–77.7%) were higher than those of AI-TPS (59.6%).

Comparison Between the Pathologists' Tumor Proportion Score and Artificial Intelligence–Assisted Tumor Proportion Score Using Mixed-Model Analysis

For a total of 584 cases, the OR for AI-assisted TPS methods 1, 2, 3, and 4 was 0.94 to 1.07 and not statistically significant. Regarding the subclass of adenocarcinomas, the OR for the 4 methods was 0.73 to 1.05 and not statistically significant. With respect to the subclass of squamous cell carcinomas, the OR for the 4 methods was 0.96 to 1.23 and not statistically significant. Considering the subclasses of discordant cases, the ORs of AI-assisted TPS methods 1, 2, 3, and 4 were 1.28 (95% CI, 1.06–1.54; $P = .012$), 1.29 (95% CI, 1.06–1.55; $P = .010$), 1.28 (95% CI, 1.06–1.54; $P = .012$), and 1.29 (95% CI, 1.06–1.55; $P = .010$), respectively, which were statistically significant. The results are summarized in [Table 2](#). Regarding the discordant cases, the OR for each AI-assisted TPS method compared with the others was 0.99 to 1.01 and not statistically significant ([Table 3](#)).

Study 4: Assessment of the Artificial Intelligence–Assisted System Using Surgical Specimens

Of the 47 cores, 23, 18, and 6 cases were classified as TPS <1% cases; ≥1% and <50% cases; and ≥50% cases, respectively. The confusion matrix for the AI-TPS is shown in [Supplementary Figure S2A](#). Each value in the matrix is presented as a percentage of the total number of true labels. Of the 47 cases, 44.7% (21/47) were matched for GT. The confusion matrix for discordant cases is shown in [Supplementary Figure S2B](#). Of the 24 discordant cases, 62.5% (15/24) were matched. The detailed

Table 2

Mixed-model analysis

Mode	All cases (n = 584)		ADCs (n = 278)		SCCs (n = 292)		Discordant cases (n = 332)	
	OR (95% CI)	P value ^a	OR (95% CI)	P value ^a	OR (95% CI)	P value ^a	OR (95% CI)	P value ^a
Path-TPS	1		1		1		1	
AI-assisted TPS 1	1.07 (0.90-1.28)	.449	1.05 (0.81-1.36)	.719	1.13 (0.89-1.45)	.323	1.28 (1.06-1.54)	.012
AI-assisted TPS 2	1.05 (0.88-1.25)	.578	0.91 (0.71-1.17)	.474	1.23 (0.96-1.58)	.101	1.29 (1.06-1.55)	.010
AI-assisted TPS 3	0.94 (0.79-1.12)	.520	0.97 (0.75-1.25)	.819	0.96 (0.76-1.22)	.759	1.28 (1.06-1.54)	.012
AI-assisted TPS 4	0.94 (0.79-1.12)	.478	0.73 (0.57-0.94)	.015	1.20 (0.94-1.54)	.146	1.29 (1.06-1.55)	.010

ADC, adenocarcinoma; AI, artificial intelligence; AI-assisted TPS, scoring in the AI-assisted system; CI, confidence interval; OR, odds ratio; Path-TPS, pathologists' scoring without AI; SCC, squamous cell carcinoma; TPS, tumor proportion score.

^a Calculated using the type III test.

agreements of AI-TPS, Path-TPS, and AI-assisted TPS for surgical specimens are presented in [Supplementary Table S5](#). In all 47 cases, the agreements of Path-TPS of all pathologists (66.0%-91.5%) were higher than those of AI-TPS (44.7%). AI-assisted TPS of 2 of the 3 pathologists (P1, 87.2%; P3, 78.7%) was higher than Path-TPS (P1, 83.0%; P3, 66.0%). A similar trend was observed in the discordant cases; agreements of AI-TPS, Path-TPS, and AI-assisted TPS were 62.5%, 33.3% to 83.3%, and 66.7% to 79.2%, respectively.

Discussion

We developed a DL model to calculate the TPS of the PD-L1 22C3 assay and evaluated whether and how the DL model could help pathologists determine the TPS. We assessed the OR of TPS using the AI-assisted system and compared it with TPS by pathologists alone using a mixed model. In 584 cases, there were no significant differences between the AI-assisted system and pathologist scoring, whereas in discordant cases, in which the pathologists' judgment was inconsistent, all 4 methods of the AI-assisted system showed higher OR. No significant differences were observed between the second- and concurrent-reading modes of the AI-assisted system. Since the concurrent-reading mode may be less burdensome for pathologists in actual clinical practice, the AI-assisted system in the concurrent-reading mode could have the advantage of reducing the burden on pathologists for the determination of PD-L1 IHC assays and improving accuracy in difficult cases.

ICIs targeting PD-L1 improve the survival of patients with NSCLCs; the PD-L1 IHC assay (22C3) has been approved as a companion diagnostic tool. Guidelines for 22C3 staining assessment are subjective, and the scoring variability among pathologists can be problematic. As revealed in several studies,¹²⁻¹⁴ we first evaluated Fleiss' κ coefficients for several cutoff points. In the literature, the κ coefficients for a 1%

cutoff 2-tiered assessment and a 50% cutoff 2-tiered assessment were 0.537 to 0.633 and 0.58 to 0.834, respectively¹²⁻¹⁴ ([Supplementary Table S3](#)). In this study, the values were 0.619 and 0.794, respectively, which are almost identical to the results of previous studies. Furthermore, the interobserver variability for the 3-tiered assessment (<1%, $\geq$ 1% and <50%, and $\geq$ 50%) was a Fleiss' κ coefficient of 0.581. It was considered moderate, despite being lower than the 1% or 50% cutoff 2-tiered assessment. We suggest that interobserver variability could be problematic when scoring PD-L1 IHC.

To achieve an objective, repeatable, and accurate PD-L1 evaluation method, several researchers have developed DL models for the PD-L1 IHC (22C3) assay for NSCLC and concluded that the automated PD-L1 scoring system showed accuracy similar to that of pathologists: Liu et al²¹ (accuracy, 74.51%), Cheng et al²⁵ (linear correlation coefficient, 0.819-0.878), Pan et al²² (intraclass correlation coefficient, 0.737), Taylor et al²³ (Lin's concordance correlation coefficient, 0.80), and Hondelink et al²⁶ (concordance, 79%). In our study, the F1 score and overall accuracy of the DL model were 0.762 and 67.6%, respectively. Although evaluations of the DL model in previous reports were not uniform, we suggest that the developed DL model demonstrated sufficient accuracy.

In the field of radiology, several authors have reported comparative studies between CAD in the second- and concurrent-reading modes.²⁸⁻³⁰ Studies by Beyer et al³⁰ and Foti et al²⁹ suggested the superiority of the second-reading mode over the concurrent-reading mode. However, Matsumoto et al²⁸ showed a conflicting result that the detection performance of radiologists in these 2 modes was not significantly different. Recently, Wu et al²⁴ assessed an AI-assisted system and reported that the intraclass correlation coefficient of untrained pathologists with AI assistance was higher than that of trained pathologists without AI assistance. To the best of our knowledge, no comparative studies have focused on the second- and concurrent-reading modes of an AI-assisted system in determining PD-L1 TPS. Our study demonstrated that the mixed-model analysis for 584 cases and for cases of adenocarcinomas and squamous cell carcinomas revealed no statistically significant difference between pathologist scoring without and with AI, even in both second- and concurrent-reading modes. However, there was a statistically significant difference between pathologist scoring without AI and all methods of the AI-assisted system when discordant cases were evaluated. Further analysis of discordant cases revealed that there were no significant differences among each method of the AI-assisted system: (1) second-reading mode referring to the overlay image, (2) second-reading mode without the overlay image, (3) concurrent-

Table 3

Mixed-model analysis

Pair	Odds ratio (95% CI)	P value ^a
AI-assisted TPS 1 vs AI-assisted TPS 4	0.99 (0.79-1.25)	.953
AI-assisted TPS 2 vs AI-assisted TPS 4	1.00 (0.79-1.26)	1.000
AI-assisted TPS 3 vs AI-assisted TPS 4	0.99 (0.79-1.25)	.953
AI-assisted TPS 1 vs AI-assisted TPS 3	1.00 (0.79-1.26)	1.000
AI-assisted TPS 2 vs AI-assisted TPS 3	1.01 (0.80-1.27)	.953
AI-assisted TPS 1 vs AI-assisted TPS 2	0.99 (0.79-1.25)	.953

x vs y means the odds ratio of x compared with the odds ratio of y.

AI, artificial intelligence; AI-assisted TPS, TPS in the AI-assisted system; CI, confidence interval; OR, odds ratio; TPS, tumor proportion score.

^a Calculated using the type III test.

reading mode referring to the overlay image, and (4) concurrent-reading mode without the overlay image. Our subgroup analysis of discordant cases showed characteristics similar to those reported previously, which are as follows^{12,38}: weak membranous staining, concomitant cytoplasmic staining, and the presence of inflammatory cells that are strongly PD-L1 positive. The usefulness of the AI-assisted system has been suggested for cases in which pathologists had difficulty in determining the PD-L1 TPS.

We performed further assessment of the AI-assisted system using a small number of surgical specimens. Among the 47 cases, AI-assisted TPS of 2 of the 3 pathologists was higher than Path-TPS. Regarding discordant cases, AI-assisted TPS was also higher than Path-TPS. A similar trend to the TMA study was observed, suggesting that the AI-assisted system will help pathologists score TPS in actual clinical practice.

This study has some limitations. First, this was a single-center study with a limited number of pathologists. We analyzed the largest number of NSCLC cases. However, further multicenter studies with more pathologists are required to validate the AI-assisted PD-L1 scoring system. Second, there may be room for improvement in the accuracy of the developed DL model. More accurate scoring using the DL model may improve the accuracy of the AI-assisted system. The accuracy of the DL model in PD-L1 IHC scoring is considered limited because it is evaluated using semi-quantitative GT by pathologists. However, further studies on AI-assisted systems with more accurate DL models are required. Third, the heterogeneity of PD-L1 expression in NSCLC could be problematic in the study using TMA samples. However, a previous study revealed that the sensitivity and specificity of scoring using a single TMA to predict PD-L1 staining on the whole tumor are relatively high: 85.1% to 100% and 97.9% to 100%, respectively, depending on the threshold (>1%, >10%, and >50%).³⁹ We selected an appropriate and representative region without necrosis when developing the TMA. Further assessment with a greater number of surgical specimens and statistical analysis will be needed. Finally, we performed repeated PD-L1 scoring on the same data sets and discussed the learning effects. Each sample of 584 cases was scored in a random order, and an appropriate washout period was taken for each scoring. Therefore, the learning effect was minimized or eliminated.

In conclusion, in total 584 cases, there were no significant differences between the AI-assisted system and pathologist scoring, whereas, in discordant cases, all the 4 methods of the AI-assisted system showed higher ORs. Regarding the methods of the AI-assisted system, there were no significant differences between second- and concurrent-reading modes. This study suggests the usefulness of the AI-assisted system for cases in which pathologists might have difficulty determining the PD-L1 TPS. Further future studies are needed to validate the AI-assisted PD-L1 scoring system.

Author Contributions

H.I., K.T., and A.Y. designed the study, analyzed the clinicopathologic data, and performed the immunohistochemical assessment of PD-L1. M.R.-K., N.N., S.S., and A.Y. selected cases and constructed the tissue microarray. H.I. annotated the whole-slide images confirmed by A.Y. T.S., K.N., and R.B. constructed the deep learning model and helped with the data analysis. A.N. and S.M. performed the mixed-model analysis. M.H., T.M., and H.D. provided the cases and extracted the clinical data. All authors

reviewed the paper and approved the final draft of the manuscript.

Data Availability

The data sets used and analyzed in the study are available from the corresponding author upon reasonable request.

Funding

This work was partly supported by the Japan Society for the Promotion of Science, KAKENHI (grant number: 17K08740).

Declaration of Competing Interest

None of the authors have any conflicts of interest to disclose.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of Kyoto University, Kyoto, Japan (reference number: R1876-1).

Acknowledgments

The authors thank Editage (www.editage.jp) for English language editing.

Supplementary Material

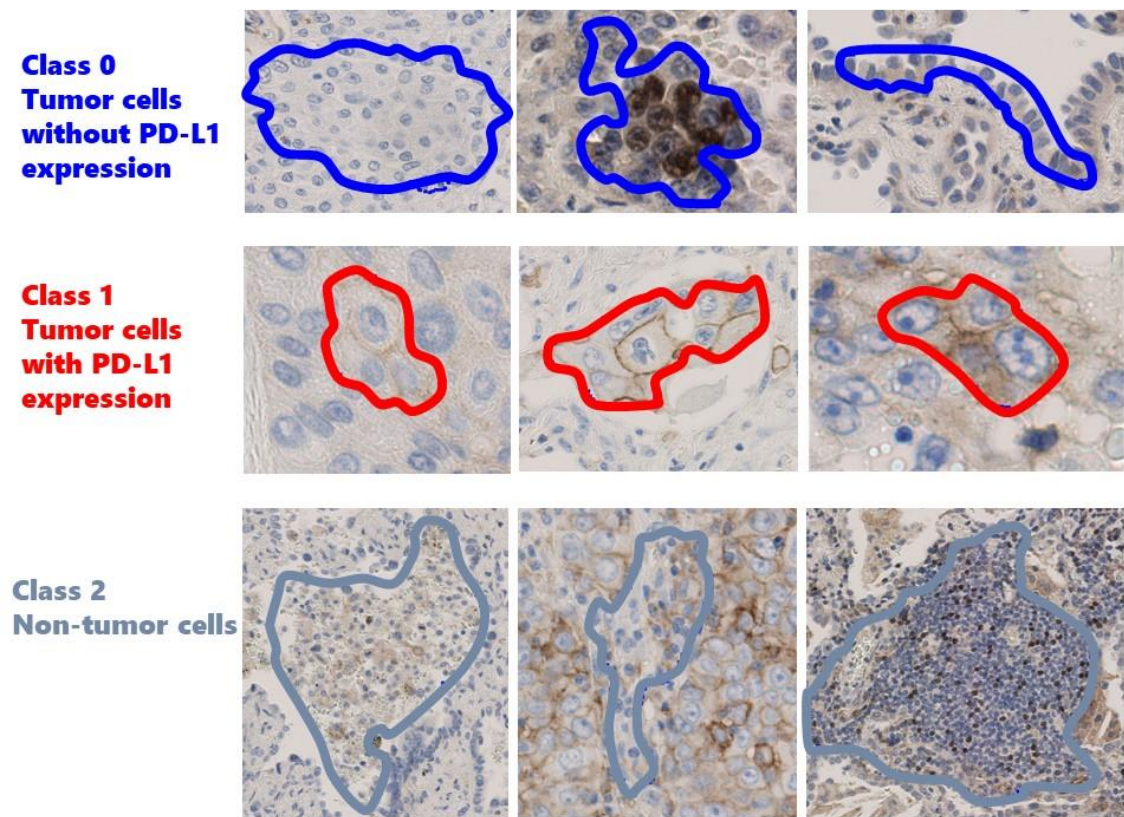
The online version contains supplementary material available at <https://doi.org/10.1016/j.modpat.2024.100485>

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71(3):209–249. <https://doi.org/10.3322/caac.21660>
2. Topalian SL, Hodi FS, Brahmer JR, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med.* 2012;366(26):2443–2454. <https://doi.org/10.1056/NEJMoa1200690>
3. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. *Nat Rev Cancer.* 2012;12(4):252–264. <https://doi.org/10.1038/nrc3239>
4. Jain KK. Personalized immuno-oncology. *Med Princ Pract.* 2021;30(1):1–16. <https://doi.org/10.1159/000511107>
5. Garon EB, Rizvi NA, Hui R, et al. Pembrolizumab for the treatment of non-small-cell lung cancer. *N Engl J Med.* 2015;372(21):2018–2028. <https://doi.org/10.1056/NEJMoa1501824>
6. Reck M, Rodríguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N Engl J Med.* 2016;375(19):1823–1833. <https://doi.org/10.1056/NEJMoa1606774>
7. Lee CK, Man J, Lord S, et al. Clinical and molecular characteristics associated with survival among patients treated with checkpoint inhibitors for advanced non-small cell lung carcinoma: a systematic review and meta-analysis. *JAMA Oncol.* 2018;4(2):210–216. <https://doi.org/10.1001/jamaoncol.2017.4427>
8. Herbst R, Jassem J, Abogunrin S, et al. A network meta-analysis of cancer immunotherapies versus chemotherapy for first-line treatment of patients with non-small cell lung cancer and high programmed death-ligand 1 expression. *Front Oncol.* 2021;11:676732. <https://doi.org/10.3389/fonc.2021.676732>
9. Sul J, Blumenthal GM, Jiang X, He K, Keegan P, Pazdur R. 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(5):643–650. <https://doi.org/10.1634/theoncologist.2015-0498>

10. Dako ApS. PD-L1 IHC 22C3 pharmDx Interpretation Manual: NSCLC. Dako; 2018.
11. Brunnström H, Johansson A, Westbom-Fremer S, et al. PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher than assay variability. *Mod Pathol*. 2017;30(10):1411–1421. <https://doi.org/10.1038/modpathol.2017.59>
12. Cooper WA, Russell PA, Cherian M, et al. Intra- and interobserver reproducibility assessment of PD-L1 biomarker in non-small cell lung cancer. *Clin Cancer Res*. 2017;23(16):4569–4577. <https://doi.org/10.1158/1078-0432.CCR-17-0151>
13. Rimm DL, Han G, Taube JM, et al. A prospective, multi-institutional, pathologist-based assessment of 4 immunohistochemistry assays for PD-L1 expression in non-small cell lung cancer. *JAMA Oncol*. 2017;3(8):1051–1058. <https://doi.org/10.1001/jamaoncol.2017.0013>
14. Chang S, Park HK, Choi YL, Jang SJ. Interobserver reproducibility of PD-L1 biomarker in non-small cell lung cancer: a multi-institutional study by 27 pathologists. *J Pathol Transl Med*. 2019;53(6):347–353. <https://doi.org/10.4132/jptm.2019.09.29>
15. Beijndorf BE, Litjens G, van der Laak JA. Machine learning compared with pathologist assessment-reply. *JAMA*. 2018;319(16):1726. <https://doi.org/10.1001/jama.2018.1478>
16. Mahmood T, Arsalan M, Owais M, Lee MB, Park KR. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. *J Clin Med*. 2020;9(3):749. <https://doi.org/10.3390/jcm9030749>
17. Zlotta AR, Sadeghian A. Re: artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Eur Urol*. 2020;78(2):290–291. <https://doi.org/10.1016/j.eururo.2020.04.010>
18. Chen PC, Gadeipalli K, MacDonald R, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat Med*. 2019;25(9):1453–1457. <https://doi.org/10.1038/s41591-019-0539-7>
19. Saltz J, Gupta R, Hou L, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep*. 2018;23(1):181–193.e7. <https://doi.org/10.1016/j.celrep.2018.03.086>
20. 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. <https://doi.org/10.1038/s41598-018-35501-5>
21. 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. <https://doi.org/10.1038/s41598-021-95372-1>
22. 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. <https://doi.org/10.1186/s12967-021-02898-z>
23. Taylor CR, Jadhav AP, Gholap A, et al. A multi-institutional study to evaluate automated whole slide scoring of immunohistochemistry for assessment of programmed death-ligand 1 (PD-L1) expression in non-small cell lung cancer. *Appl Immunohistochem Mol Morphol*. 2019;27(4):263–269. <https://doi.org/10.1097/PAI.0000000000000737>
24. 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. <https://doi.org/10.1038/s41379-021-00904-9>
25. 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. <https://doi.org/10.3389/fimmu.2022.893198>
26. Hondelink LM, Hüyük 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. <https://doi.org/10.1111/his.14571>
27. Naso JR, Povshedna T, Wang G, et al. Automated PD-L1 scoring for non-small cell lung carcinoma using open-source software. *Pathol Oncol Res*. 2021;27:609717. <https://doi.org/10.3389/pore.2021.609717>
28. Matsumoto S, Ohno Y, Aoki T, et al. Computer-aided detection of lung nodules on multidetector CT in concurrent-reader and second-reader modes: a comparative study. *Eur J Radiol*. 2013;82(8):1332–1337. <https://doi.org/10.1016/j.ejrad.2013.02.005>
29. Foti G, Faccioli N, D'Onofrio M, Contro A, Milazzo T, Pozzi Mucelli R. Evaluation of a method of computer-aided detection (CAD) of pulmonary nodules with computed tomography. *Radiol Med*. 2010;115(6):950–961. <https://doi.org/10.1007/s11547-010-0556-6>
30. Beyer F, Zierott L, Fallenberg EM, et al. Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader. *Eur Radiol*. 2007;17(11):2941–2947. <https://doi.org/10.1007/s00330-007-0667-1>
31. Amin MB, American Joint Committee on Cancer, American Cancer Society. 36. Lung. In: Amin MB, Edge Stephen B, Gress Donna M, et al., eds. *AJCC Cancer Staging Manual*. 8th edition. American Joint Committee on Cancer, Springer; 2017:447.
32. WHO Classification of Tumours Editorial Board. Thoracic tumours. *WHO Classification of Tumours Series*. 5th ed; vol 5. International Agency for Research on Cancer; 2021.
33. Tsao MS, Kerr KM, Dacic S, Yatabe Y, Hirsch FR, eds. *IASLC Atlas of PD-L1 Immunohistochemistry Testing in Lung Cancer*. IASLC; 2017.
34. Multi-class cell detection using modified self-attention. In: Sugimoto T, Ito H, Teramoto Y, Yoshizawa A, Bise R, eds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2022:1854–1862. <https://doi.org/10.1109/CVPRW56347.2022.00202>
35. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H, eds. *Proceedings of the European Conference on Computer Vision*. ECCV; 2018: 801–818.
36. Multilabel soft margin loss. The PyTorch Foundation. The Linux Foundation. Accessed February 8, 2024. <https://pytorch.org/docs/stable/generated/torch.nn.MultiLabelSoftMarginLoss.html>
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. <https://doi.org/10.2307/2529310>
38. Humphries MP, McQuaid S, Craig SG, et al. Critical appraisal of programmed death ligand 1 reflex diagnostic testing: current standards and future opportunities. *J Thorac Oncol*. 2019;14(1):45–53. <https://doi.org/10.1016/j.jtho.2018.09.025>
39. Gniadek TJ, Li QK, Tully E, Chatterjee S, Nimmagadda S, Gabrielson E. Heterogeneous expression of PD-L1 in pulmonary squamous cell carcinoma and adenocarcinoma: implications for assessment by small biopsy. *Mod Pathol*. 2017;30(4):530–538. <https://doi.org/10.1038/modpathol.2016.213>

Supplementary Fig. S1.



Supplementary Fig. S2.

A		Total (n = 47)		
Ground truth	$\geq 50\%$	2 (8.7%)	21 (91.3%)	0 (0.0%)
	$\geq 1\%$ and $< 50\%$	1 (5.6%)	17 (94.4%)	0 (0.0%)
	$< 1\%$	0 (0.0%)	4 (66.7%)	2 (33.3%)
	AI-TPS			

B		Discordant cases (n = 24)		
Ground truth	$\geq 50\%$	1 (11.1%)	8 (88.9%)	0 (0.0%)
	$\geq 1\%$ and $< 50\%$	0 (0.0%)	14 (100.0%)	0 (0.0%)
	$< 1\%$	0 (0.0%)	1 (100.0%)	0 (0.0%)
	AI-TPS			

Supplementary Figure legends

Supplementary Figure S1.

Examples of annotations. Top images are representative of class 0 (tumor cells without PD-L1 expression); middle, class 1 (tumor cells with PD-L1 expression); and bottom, class 2 (non-tumor cells). The enclosed region of the images shows the annotated region for each class.

Supplementary Figure S2.

Confusion matrices of AI-TPS for surgical specimens. A: confusion matrix for 47 cases; B: discordant cases (24 cases).

Supplementary Table S1: Threefold cross-validation

	Positive tumor area	Negative tumor area	Non-tumor area
Fold 1	12142502	9205227	93296362
Fold 2	7571491	9189820	96565959
Fold 3	6931828	9163142	93973548

Supplementary Table S2: Threefold cross-validation

Test set	mPrecision	mRecall	F1
Fold 1	0.710	0.888	0.789
Fold 2	0.689	0.892	0.777
Fold 3	0.660	0.790	0.719
Average	0.686	0.857	0.762

Supplementary Table S3: Summary of interobserver variability (Fleiss' kappa coefficient)

	1% and 50% cut-off	1% cut-off	50% cut-off
Our study*	0.581	0.619	0.794
Rimm	NA	0.537	0.749
Cooper	NA	0.68	0.58
Chang	NA	0.633	0.834

Abbreviation: NA, not available

*Shown as the average of three scores of pathologist's scoring without AI

Supplementary Table S4: Agreement (overall accuracy)

	All cases (N = 584)			ADCs (n = 278)			SCCs (n = 292)			Discordant cases (n = 332)		
AI-TPS	67.6			66.2			68.8			59.6		
	P1	P2	P3	P1	P2	P3	P1	P2	P3	P1	P2	P3
Path-TPS*	83.6	79.8	75.8	84.1	81.8	76.4	82.9	77.5	74.7	71.1	64.5	57.4
AI-assisted TPS 1)	87.2	80.7	73.3	89.2	80.6	73.7	85.6	80.1	72.9	77.7	66.0	59.9
AI-assisted TPS 2)	85.1	79.6	75.9	84.9	79.5	75.2	85.3	79.1	76.7	75.0	64.2	64.8
AI-assisted TPS 3)	80.1	81.2	76.2	82.7	81.7	77.0	78.1	80.1	75.7	70.5	66.9	66.3
AI-assisted TPS 4)	85.1	78.3	74.0	83.1	76.3	73.7	86.6	79.5	74.3	77.1	62.3	64.5

Abbreviations: ADC, adenocarcinoma; SCC, squamous cell carcinoma; TPS, tumor proportion score; AI-TPS, scoring by AI; Path-TPS, pathologist's scoring without AI; AI-assisted TPS, scoring in AI-assisted system; P1–3, pathologist 1–3

*Shown as the average of three scores

Supplementary Table S5: Agreement (overall accuracy) for surgical specimens

	All cases (N = 47)			Discordant cases (n = 24)		
AI-TPS	44.7			62.5		
	P1	P2	P3	P1	P2	P3
Path-TPS	83.0	91.5	66.0	66.7	83.3	33.3
AI-assisted TPS	87.2	83.0	78.7	79.2	66.7	66.7

Abbreviations: TPS, tumor proportion score; AI-TPS, scoring by AI; Path-TPS, pathologist's scoring without AI; AI-assisted TPS, scoring in AI-assisted system; P1–3, pathologist 1–3