Non-invasive measurement of PD-L1 status and prediction of immunotherapy response using deep learning of PET/CT images

Wei Mu,1 Lei Jiang,2 Yu Shi,3 Ilke Tunali,1 Jhanelle E Gray,4 Evangelia Katsoulakis,5 Jie Tian,6,7 Robert J Gillies,1 Matthew B Schabath6,4,8

ABSTRACT

Background Currently, only a fraction of patients with non-small cell lung cancer (NSCLC) treated with immune checkpoint inhibitors (ICIs) experience a durable clinical benefit (DCB). According to NCCN guidelines, Programmed death-ligand 1 (PD-L1) expression status determined by immunohistochemistry (IHC) is the only clinically approved companion biomarker to trigger the use of ICI therapy. Based on prior work showing a relationship between quantitative imaging and gene expression, we hypothesize that quantitative imaging (radiomics) can provide an alternative surrogate for PD-L1 expression status in clinical decision support.

Methods 18F-FDG-PET/CT images and clinical data were curated from 697 patients with NSCLC from three institutions and these were analyzed using a small-residual-convolutional-network (SResCNN) to develop a deeply learned score (DLS) to predict the PD-L1 expression status. This developed model was further used to predict DCB, progression-free survival (PFS), and overall survival (OS) in two retrospective and one prospective test cohorts of ICI-treated patients with advanced stage NSCLC.

Results The PD-L1 DLS significantly discriminated between PD-L1 positive and negative patients (area under receiver operating characteristics curve ≥0.82 in the training, validation, and two external test cohorts). Importantly, the DLS was indistinguishable from IHC-derived PD-L1 status in predicting PFS and OS, suggesting the utility of DLS as a surrogate for IHC. A score generated by combining the DLS with clinical characteristics was able to accurately (C-indexes of 0.70–0.87) predict DCB, PFS, and OS in retrospective training, prospective testing and external validation cohorts.

Conclusion Hence, we propose DLS as a surrogate or substitute for IHC-determined PD-L1 measurement to guide individual pretherapy decisions pending in larger prospective trials.

INTRODUCTION

The emergence of immune checkpoint inhibitors (ICIs) has revolutionized cancer treatment and improved long-term survival among some patients with advanced stage non-small cell lung cancer (NSCLC), but durable clinical benefit (DCB) is only observed in 20%–50% patients.1,2 Because of the complexity and heterogeneity of response, NCCN guidelines recommend treatment based on expression of the checkpoint target, programmed death-ligand 1 (PD-L1), determined by immunohistochemistry (IHC).3 Early studies showed that PD-L1 positivity is associated with significantly higher objective response rate, longer progression-free survival (PFS), and longer overall survival (OS).3,4 However, measuring PD-L1 by IHC requires surgical or biopsied tumor specimens, which are collected through invasive procedures and associated with risk of morbidities.5 Therefore, an alternative non-invasive method of measuring PD-L1 status would have important implications for clinical decision support, especially when tissues are not available or when the IHC fails.5

Radiomic analyses of quantitative image features based on shape, size, voxel intensity, and texture are strongly associated with gene and protein expression in NSCLC.7,8 Signatures are typically extracted from the intratumoral region, but it is becoming increasingly appreciated that the peritumoral region,9 encompassing the tumor-stroma interface, is also informative in predictive models, likely because this region contains information on immune infiltration and stromal inflammation. Intratumoral and peritumoral immune-cell infiltration is necessary for inducing an immunotherapy response. Immune infiltration is associated with expression of cell checkpoint markers including PD-L1,10 which is significantly correlated with metabolic rate,11 GLUT-1 expression,12 pAKT levels,13 hypoxia, and acidosis.14 These observations suggest that PD-L1 expression might be tractable by radiomics analyses of Fluorine 18 (18F)-fluorodeoxyglucose (FDG)-Positron emission tomography (PET) scans. As a consequence,
others have investigated the relationship between FDG-PET and PD-L1 status in NSCLC, but these analyses were limited to a few statistical associations.15 16 Our previous study demonstrated the utility of deep learning methods using intratumoral and peritumoral radiomics from PET/CT images to predict epidermal growth factor receptor (EGFR) mutation status, which could be used to support the treatment decisions for EGFR-TKI and other therapies, including ICI, which is generally more effective in EGFR wild-type cancers.17

In the current work, we utilized machine learning to develop and validate a deeply learned score (DLS) to measure PD-L1 expression status non-invasively using pretreatment ¹⁸F-FDG PET/CT images of a retrospective cohort accrued from Shanghai Pulmonary Hospital (SPH). Then, to validate the DLS in accordance with the FDA guidance document for the Clinical Evaluation of Software as a Medical Device (SaMD),18 clinical association analysis for scientific validity, analytic validation analysis for accuracy and reliability, and clinical validation analysis for identifying patients most likely to benefit from ICI treatment were performed using external test cohorts from the H Lee Moffitt Cancer Center and Research Institute (MCC) with both PD-L1 status and clinical follow-up information (ie, MCC PD-L1 cohort) or only with clinical follow-up information (ie, MCC ICI-treated retrospective and prospective cohorts). To determine the potential application for accurate quantitative prognostic prediction, we developed DCB, PFS, and OS prediction models with the derived DLS using the MCC ICI-treated retrospective cohort. The models for all three endpoints were independently tested with the MCC ICI-treated prospective cohort. Finally, an external ICI-treated cohort from a third institution, James A Haley Veterans’ Administration (VA), was used to blindly validate the models mentioned above (details shown in figure 1 and online supplemental figure S1).

![Diagram of study design]

**Figure 1** Study design, which contains three main phases. First, the SPH data comprised PD-L1 expression data and the corresponding imaging data was used to train and validate the deeply learned score (DLS). Then, according to FDA SaMD guideline, the DLS was evaluated through the clinical association and analytic validation on the two cohorts (MCC PD-L1 data and external VA PD-L1 data), which had both PD-L1 expression data and clinical follow-up information, as well as the clinical validation on three other cohorts (MCC ICI-treated retrospective, prospective, and external VA ICI-treated cohorts), which had clinical follow-up information. Third, in order to further test the application of DLS in guiding treatment, the well-validated DLS was utilized to develop prognosis prediction models with the MCC ICI-treated retrospective cohort, which was tested with MCC ICI-treated prospective and external VA ICI-treated cohorts. DCB, durable clinical benefit; ICI, immune checkpoint inhibitor; MCC, H Lee Moffitt Cancer Center and Research Institute; ROI, region of interest; SPH, Shanghai Pulmonary Hospital; VA, James A Haley Veterans’ Administration; EGFR, epidermal growth factor receptor; PD-L1, programmed death-ligand 1; FDA, Food and Drug Administration; PET/CT, positron emission tomography/computed tomography.
Study population
In this multi-institutional study, five cohorts of patients were first accrued from two institutions: SPH, Shanghai, China, and MCC, Tampa, Florida. The detailed inclusion and exclusion criteria are provided in online supplemental figure S1 and methods S1. Among these, the SPH retrospective cohort, which was split into training (N=284) and validation (N=116) cohorts randomly by 71%–29%, and the retrospective MCC cohort with PD-L1 status (N=85) were used for training, validating, and testing the DLS to measure PD-L1 expression status non-invasively; one ICI-treated retrospective cohort (N=128) and one ICI-treated prospective cohort (N=49) were used to validate the prognostic value of the DLSs and investigate the association of the DLS and clinical characteristics on the clinical outcomes. Additionally, a sixth cohort (N=35) from the third institution, VA, Tampa, Florida, was curated as an external validation of the DLS and the prognostic models.

The progression of the distinct ICI-treated cohorts used to investigate the association of the DLS and clinical characteristics with the clinical outcome including DCB (PFS >6 months9), PFS, and OS, were defined using Response Evaluation Criteria in Solid Tumors (RECIST V.1.1).20 Detailed acquisition parameters for the 18F-FDG PET/CT imaging for each cohort are presented in online supplemental table S1. All PET images were converted into standardized uptake value (SUV) units by normalizing the activity concentration to the dosage of 18F-FDG injected and the patient’s body weight after decay correction.

18F-FDG PET/CT imaging
Detailed acquisition parameters for the 18F-FDG PET/CT imaging for each cohort are presented in online supplemental figure S2. For each patient, only the primary tumor was analyzed. A square or an irregular-shaped box, which was close to the boundary of the tumor, was delineated manually in the aligned PET and CT images of the SPH cohort using ITK-SNAP software by experienced nuclear medicine radiologist (LJ). After dilation of the smallest square mask (SSM) including the selected region with a square of size 20 mm and resized to the size of 64×64 pixels using cubic spline interpolation, the PET region of interest (ROI) and CT ROI with the entire tumor and its peripheral region included were automatically generated at the same size (online supplemental figure S3). To reduce the effect of the difference between the central slice and peripheral slices, the area of each SSM within each patient was calculated, and only the SSMS with area larger than 30% of the maximum value were used to generate valid ROIs, which further constructed a three-channel hyper-images together with their fusion images (alpha-blending fusion, α=1, online supplemental figure S4). During the training of the model, 14,011 training hyper-images (6722 were PD-L1 positive and 7289 were PD-L1 negative) and 5291 validation hyper-images (2513 were PD-L1 positive and 2778 were PD-L1 negative) were used as the input images, the PD-L1 expression status (positive=1 or negative=0) was used as the label. After training, a DLS representing the PD-L1 positivity status was generated after a sequential activation of convolution and pooling layers. To develop a robust measurement, the average DLSs of all valid slices including tumor tissue fed into the SResCNN model with equal weight were regarded as the PD-L1 positive probability of the tumor. Details of the building, training, optimization, and application methods were provided in online supplemental methods S3. The implementation of this model used the Keras toolkit and Python 3.5. The same pipeline (available at https://doi.org/10.5281/zenodo.4731166) was performed by an experienced radiologist (YS) on the three MCC cohorts and external VA cohort to obtain the DLS based on the guideline. Given there were minor differences between the different radiologists in selecting the ROIs, ROIs within the SPH validation cohort were also selected by YS again to validate the reproducibility of DLS. Regarding the importance of the hyper-image constructed with different modalities, similar SResCNN models using only PET or CT images were also trained.

PD-L1 expression by IHC
The detailed information of IHC staining for PD-L1 expression is provided in online supplemental methods S2. For both SPH and MCC PD-L1 cohort, the platform of Dako Link 48 and the antibody of Dako 22C3 were used for PD-L1 staining to quantify the presence of PD-L1. The level of PD-L1 expression was presented as a tumor proportion score (TPS), which is the percentage of viable tumor cells showing membrane PD-L1 staining relative to all viable tumor cells and is given as 0%, 1%–49%, and ≥50%, and PD-L1 positivity was defined as ≥1% of TPS.21 22 To compensate for reader bias, all the staining results were reviewed and analyzed by two experienced pathologists who were blinded to each other’s scores and unaware of the patients’ clinical information. When there was discrepancy, the two pathologists would have a mutual discussion to reach a consensus.

Development of the DLS
The architecture of the small-residual-convolutional-network (SResCNN) model used for measuring PD-L1 expression non-invasively is presented in online supplemental figure S2. For each patient, only the primary tumor was analyzed. A square or an irregular-shaped box, which was close to the boundary of the tumor, was delineated manually in the aligned PET and CT images of the SPH cohort using ITK-SNAP software by experienced nuclear medicine radiologist (LJ). After dilation of the smallest square mask (SSM) including the selected region with a square of size 20 mm and resized to the size of 64×64 pixels using cubic spline interpolation, the PET region of interest (ROI) and CT ROI with the entire tumor and its peripheral region included were automatically generated at the same size (online supplemental figure S3). To reduce the effect of the difference between the central slice and peripheral slices, the area of each SSM within each patient was calculated, and only the SSMS with area larger than 30% of the maximum value were used to generate valid ROIs, which further constructed a three-channel hyper-images together with their fusion images (alpha-blending fusion, α=1, online supplemental figure S4). During the training of the model, 14,011 training hyper-images (6722 were PD-L1 positive and 7289 were PD-L1 negative) and 5291 validation hyper-images (2513 were PD-L1 positive and 2778 were PD-L1 negative) were used as the input images, the PD-L1 expression status (positive=1 or negative=0) was used as the label. After training, a DLS representing the PD-L1 positivity status was generated after a sequential activation of convolution and pooling layers. To develop a robust measurement, the average DLSs of all valid slices including tumor tissue fed into the SResCNN model with equal weight were regarded as the final PD-L1 positive probability of the tumor. Details of the building, training, optimization, and application methods were provided in online supplemental methods S3. The implementation of this model used the Keras toolkit and Python 3.5. The same pipeline (available at https://doi.org/10.5281/zenodo.4731166) was performed by an experienced radiologist (YS) on the three MCC cohorts and external VA cohort to obtain the DLS based on the guideline. Given there were minor differences between the different radiologists in selecting the ROIs, ROIs within the SPH validation cohort were also selected by YS again to validate the reproducibility of DLS. Regarding the importance of the hyper-image constructed with different modalities, similar SResCNN models using only PET or CT images were also trained.

Visualization of the SResCNN model
To further understand the measurement processing and explore the biological underpinnings of the deep learning features, intermediate activation layers were first visualized to assess how the network carries the information from input to output.25 Additionally, the Gradient-weighted Class Activation Mapping (Grad-CAM) was used to understand the importance of each neuron for a decision of PD-L1 positive or negative and produce a coarse localization map highlighting the important regions in the image for predicting the target concept (PD-L1 positive
or PD-L1 negative) by using the gradient information of target concept flowing into the last convolutional layer of the SResCNN model. And the reconstructed maps were named as positive and negative filter later, which were also used to evaluate the class discrimination. Besides, unsupervised hierarchical clustering was performed on the deeply learned features (ie, the output of global average pooling, N=256) to create a heatmap to show their distinguishable expression pattern among different patients. The clusters formed were based purely on the similarities and dissimilarities among the patients by the expressions of the deeply learned features.

**Statistical analysis**

The Wilcoxon signed-rank test and Fisher’s exact test were used to test the differences for continuous variables and categorical variables, respectively. The area under the receiver operating characteristics curve (AUC), accuracy (ACC), sensitivity (SEN), specificity (SPEC), and the 95% CI by the DeLong method were used to assess the ability of DLS in discriminating between positive and negative PD-L1 expression. The cutoff was established using the maximum Youden index (ie, Specificity +Sensitivity-1) in the SPH training cohort. To compare the prognostic value of DLS with that of IHC-based PD-L1 status, the difference between HRs for the DLS and PD-L1 status computed by Cox regression model was calculated and evaluated with bootstrapped 95% CI. The inter-rater agreement of DLS estimations was calculated by intra-class correlation coefficient (ICC) between two radiologists. The bootstrapped mean value and SEs of the DLSs in different cohorts were also assessed for the similarity.

The correlation between DLS and different metadata (including age, body mass index, sex, stage, smoking status, Eastern Cooperative Oncology Group (ECOG) Performance Status, and SUVmax) and molecular features (including histology, PET/CT image-based necrosis, and PD-L1 TPS) were analyzed by Spearman’s rank correlation or point-biserial correlation. The details of necrosis quantification are shown in online supplemental methods S4. Comparison of the magnitude of two correlations was performed with a software package named cccor. Given prior research has suggested that PD-L1 expression is negatively correlated with EGFR mutation status, we contend that non-invasive methods of measuring PD-L1 and/or EGFR status would have clinical translational implications. Therefore, we also investigated whether the DLS was correlated with EGFR mutation status using point-biserial correlation and whether the DLS was affected by EGFR mutation status by comparing the performance in the subgroups divided by EGFR mutation status.

In the ICI-treated cohorts, the patients were clustered into high DLS and low DLS groups with the obtained cutoff, and survival analyses were performed using Kaplan-Meier method and Cox proportional hazards model. Using the MCC ICI-treated retrospective cohort, multivariable models, including the risk factors selected in univariate analysis according to the significance, were developed for the prediction of DCB, PFS, and OS, which were tested using the MCC ICI-treated prospective cohort as well as external VA ICI-treated retrospective cohort and were evaluated with C-indices. Z test was applied to compare the differences between different models.

To rigorously assess the quality of the study design, the radiomic quality score was calculated (online supplemental methods S5 and table S2). Two-sided p values of less than 0.05 were regarded as significant, 10,000 replications were performed in bootstrap analyses, and all statistical analyses were conducted with IBM SPSS Statistics 25 (Armonk, New York, USA), R 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria), and MATLAB R2019a (Natick, Massachusetts).

**RESULTS**

**Patients characteristics**

The clinical characteristics of the patients used to train and test the non-invasive measurement of the PD-L1 status are presented in table 1 (online supplemental table S3 for external VA patients). The SPH training, SPH validation, and external MCC PD-L1 test cohorts used to train, validate, and test the SResCNN model had a prevalence of PD-L1 positivity by IHC of 29.93%, 30.17%, and 56.47%, respectively. The external VA patients had a significant higher PD-L1 positivity of 82.76% (within the 29 patients who had IHC PD-L1 expression).

The clinical characteristics of the patients used to test the clinical utility of DLS are presented in table 2. The retrospective MCC ICI-treated cohort included 128 patients with a median PFS and OS of 7.43 and 21.77 months, respectively, and 53.91% of the patients had DCB. The prospective MCC ICI-treated patients included 49 patients with a DCB rate of 65.31%, median PFS and OS of 10.50 and 17.00 months, respectively. For the external VA patients with a median PFS and OS of 8.13 and 13.10 months, 68.57% of the patients showed PD-L1 positive, and 54.29% of patients obtained DCB.

**Association between DLS, PD-L1, and metadata**

The DLS exhibited statistically significant differences between the PD-L1-positive and PD-L1-negative tumors in all three cohorts (p<0.001), and four examples are shown in figure 2 (adapted from Mu et al). The DLS was also positively correlated with the original PD-L1 TPS in both SPH (Spearman’s rho=0.60, p<0.001) and MCC PD-L1 test (Spearman’s rho=0.59, p<0.001) cohorts, which was significantly higher compared with the correlation between the SUVmax and the TPS with rho of 0.30 (p=0.001) and 0.29 (p=0.009), respectively. Using analysis of variance, the DLS was significantly different between groups with PD-L1 TPS ≤1%, 1%–49%, and ≥50% (SPH cohort: p<0.001; MCC PD-L1 test cohort: p<0.001). The least squares difference (LSD) post hoc analysis showed significantly higher values of DLS in patients with PD-L1 TPS ≥50% than TPS 1%–49% group (LSD: SPH cohort: 10.00 ± 0.00; MCC PD-L1 test cohort: 10.00 ± 0.00; SPH training cohort: 10.00 ± 0.00; SPH validation cohort: 10.00 ± 0.00).

**Association between PD-L1 expression and clinicopathological features**

The association between PD-L1 expression and clinicopathological features was also analyzed in table 1 (supplementary table S3 for external VA patients). The DLS exhibited statistically significant differences between PD-L1 positive and PD-L1 negative tumors in all three cohorts (p<0.001), and four examples are shown in figure 2 (adapted from Mu et al). The DLS was also positively correlated with the original PD-L1 TPS in both SPH (Spearman’s rho=0.60, p<0.001) and MCC PD-L1 test (Spearman’s rho=0.59, p<0.001) cohorts, which was significantly higher compared with the correlation between the SUVmax and the TPS with rho of 0.30 (p=0.001) and 0.29 (p=0.009), respectively. Using analysis of variance, the DLS was significantly different between groups with PD-L1 TPS ≤1%, 1%–49%, and ≥50% (SPH cohort: p<0.001; MCC PD-L1 test cohort: p<0.001). The least squares difference (LSD) post hoc analysis showed significantly higher values of DLS in patients with PD-L1 TPS ≥50% than TPS 1%–49% group (LSD: SPH cohort: 10.00 ± 0.00; MCC PD-L1 test cohort: 10.00 ± 0.00; SPH training cohort: 10.00 ± 0.00; SPH validation cohort: 10.00 ± 0.00).
| Characteristic                      | SPH training cohort (N=284) | SPH validation cohort (N=116) | MCC PD-L1 test cohort (N=85) |
|------------------------------------|-----------------------------|-------------------------------|------------------------------|
|                                    | PD-L1+                       | PD-L1−                        | P values                     |
|                                    | Mean±SD                      | Mean±SD                      | P values                     |
|                                    |                              |                              |                              |
| Age (years)                        | 0.41                         | 0.46                         | 0.40                         |
| Mean±SD                            | 62.71±8.78                   | 63.51±8.56                   | 68.21±9.18                   |
| Sex, No (%)                        | 0.035*                       | 0.062                        | 0.14                         |
| Male                               | 58 (68.24)                   | 108 (54.27)                  | 26 (52.08)                   |
| Female                             | 27 (31.76)                   | 91 (45.73)                   | 23 (47.92)                   |
| TNM stage                          | 0.12                         | 0.42                         | 0.23                         |
| I                                  | 43 (50.59)                   | 122 (61.31)                  | 0 (0)                        |
| II                                 | 22 (26.19)                   | 34 (17.08)                   | 2 (4.17)                     |
| III                                | 11 (13.10)                   | 31 (15.58)                   | 2 (4.17)                     |
| IV                                 | 9 (10.71)                    | 12 (6.03)                    | 42 (87.50)                   |
| Histology (baseline), No (%)       | <0.001*                      | 0.006*                       | 0.83                         |
| ADC                                | 48 (56.47)                   | 156 (78.39)                  | 26 (54.17)                   |
| SCC                                | 37 (43.53)                   | 43 (21.61)                   | 21 (56.76)                   |
| EGFR, No (%)                       | 0.020*                       | 0.28                         | 0.076                        |
| Mutation                           | 24 (28.57)                   | 87 (43.72)                   | 2 (4.17)                     |
| Wild                               | 55 (65.48)                   | 101 (50.75)                  | 5 (13.51)                    |
| ALK, No (%)                        | 0.51                         | 0.20                         | 1.00                         |
| Mutation                           | 1 (1.19)                     | 1 (0.50)                     | 2 (1.23)                     |
| Wild                               | 78 (92.86)                   | 187 (93.97)                  | 1 (2.08)                     |
| ROS1, No (%)                       | 1.00                         | NaN                          | NaN                          |
| Smoking status, No (%)             | <0.001*                      | 0.025*                       | 0.83                         |
| Never                              | 32 (37.65)                   | 118 (59.3)                   | 17 (35.42)                   |
| Former                             | 53 (62.35)                   | 81 (40.7)                    | 14 (37.84)                   |
| SUVmax                             | <0.001*                      | 0.003*                       | 0.014*                       |
| Mean±SD                            | 12.32±5.99                   | 8.42±4.83                    | 13.74±7.83                   |
| Deeply learned score (DLS)         | <0.001*                      | <0.001*                      | <0.001*                      |
| Median (IQR)                       | 0.70 (0.66–0.78)             | 0.43 (0.30–0.55)             | 0.58 (0.47–0.62)             |
| PD-L1 positivity by IHC            |                              |                              |                              |

Continued
p<0.017; MCC PD-L1 test cohort: p≤0.027) and TPS <1% group (LSD: SPH cohort: p=0.001; MCC PD-L1 test cohort: p=0.001) (details are shown in online supplemental figure S5). As such, the increased PD-L1 TPS scores correlated to the DLS.

Additionally, the DLS was positively correlated with SUVmax (Spearman’s rho=0.43, p<0.001), squamous cell carcinoma (SCC) (point-biserial rho_p=0.27, p<0.001), male sex (point-biserial rho_p=0.19, p<0.001), smoking status (point-biserial rho_p=0.20, p<0.001), and negatively correlated with EGFR status (point-biserial rho_p=-0.20, p<0.001) for the whole SPH cohort. In the MCC PD-L1 test cohort, the only positive significant correlation was with SUVmax (Spearman’s rho=0.34, p<0.001) and negative with EGFR status (point-biserial rho_p=-0.25, p=0.035). Further, multivariable linear regression (adjusted r²=0.15, F=15.31, p<0.001) showed that only SUVmax (coefficient=0.32, p=0.005) was independently associated with the DLS. Only 15% of DLS variability could be explained by the SUVmax, indicating that DLS originated mainly from other image information.

Through the visualization of the SResCNN model, the necrotic region was identified as self-learned important area for classifying PD-L1 status (figure 2A). Quantitatively, for all the SPH patients with necrotic regions, a significant correlation was observed between the necrosis-to-global volume ratio of the PET images and DLS with Spearman’s rho of 0.50 (p<0.001) (online supplemental figure S6). Further, univariable linear regression (adjusted r²=0.24, F=15.92, p<0.001) showed that the necrosis (coefficient=0.49, p<0.001) was independently associated with DLS and could explain 24% of DLS variability. Therefore, necrosis potentially played an important role in predicting PD-L1 status.

Finally, the DLS was not correlated with tumor volume, neither in the entire SPH cohort (rho=0.082, p=0.10) and nor in the MCC PD-L1 test cohort (rho=-0.066, p=0.55), which indicates the T stage has a limited effect on DLS in this study.

### Analytical validation of DLS in predicting PD-L1 status

To discriminate PD-L1-positive from PD-L1-negative expression, the DLS yielded AUCs of 0.89 (95% CI: 0.84 to 0.94; p<0.001), 0.84 (95% CI: 0.76 to 0.92; p<0.001), and 0.82 (95% CI: 0.74 to 0.89; p<0.001), accuracies of 81.69% (95% CI: 77.11% to 85.91%), 78.45% (95% CI: 71.55% to 85.30%), and 77.65% (95% CI: 69.41% to 85.88%), sensitivities of 84.71% (95% CI: 74.47% to 91.76%), 77.43% (95% CI: 57.14% to 85.71%), and 68.75% (95% CI: 55.26% to 81.25%), specificities of 80.40% (95% CI: 74.87% to 85.67%), 81.48% (95% CI: 72.84% to 88.89%), and 89.19% (95% CI: 78.38% to 97.30%) in the SPH training, SPH validation, and MCC PD-L1 test cohorts, respectively, with a cutoff value of 0.55 (figure 3 and online supplemental table S4). For the external VA PD-L1 patients, the DLS generated an AUC of 0.82 (95% CI: 0.65 to 0.98; p=0.028), accuracy of 79.31% (95% CI: 62.07% to 93.10%), sensitivity of 88.33% (93%
# Table 2  Demographic and clinical characteristics for ICI-treated patients

| Characteristic | Retrospective MCC ICI-treated patients (N=128) | Prospective MCC ICI-treated patients (N=49) |
|---------------|-----------------------------------------------|---------------------------------------------|
|               | Deeply learned score | P values | Deeply learned score | P values |
|               | All | High (N=43) | Low (N=85) | All | High (N=31) | Low (N=18) |
| Age (years)   |                                   | 0.099 | 0.20 |
|               | Mean±SD | 65.48±13.24 | 67.35±13.71 | 64.54±12.80 | 66.8±10.04 | 64.77±8.87 | 70.28±10.70 |
| BMI           |                                   | 0.43 | 0.48 |
|               | Mean±SD | 26.14±5.08 | 25.60±4.79 | 26.42±5.17 | 26.06±5.02 | 26.05±4.56 | 26.08±4.00 |
| Sex, No (%)   |                                   | 0.38 | 0.56 |
| Male          |                                   | 81 (63.28) | 24 (55.81) | 57 (67.06) | 25 (51.02) | 17 (54.84) | 8 (44.44) |
| Female        |                                   | 47 (36.72) | 19 (44.19) | 28 (32.94) | 24 (48.98) | 14 (43.75) | 10 (55.56) |
| TNM stage     |                                   | 0.63 | 0.39 |
| III           |                                   | 25 (19.53) | 11 (25.58) | 14 (16.47) | 6 (12.24) | 5 (16.13) | 1 (5.56) |
| IV            |                                   | 103 (80.47) | 32 (74.42) | 71 (83.53) | 43 (87.76) | 26 (83.87) | 17 (94.44) |
| Histology (baseline), No (%) | | 0.096 | 0.23 |
| ADC           | | 80 (62.50) | 23 (53.49) | 57 (67.06) | 28 (57.14) | 20 (64.52) | 8 (44.44) |
| SCC           | | 48 (347.50) | 20 (46.51) | 28 (32.94) | 21 (42.86) | 11 (35.48) | 10 (55.56) |
| EGFR, No (%)  |                                   | 1.00 | 1.00 |
| Mutation      | | 8 (6.25) | 2 (4.65) | 6 (7.06) | 2 (4.08) | 2 (6.45) | 0 |
| Wild          | | 85 (66.41) | 28 (65.12) | 57 (67.06) | 37 (75.51) | 23 (74.19) | 14 (77.78) |
| ALK, No (%)   |                                   | 1.00 | NaN |
| Mutation      | | 2 (1.56) | 0 | 2 (2.35) | 0 | 0 | 0 |
| Wild          | | 89 (69.53) | 27 (62.79) | 62 (72.94) | 39 (79.59) | 24 (77.42) | 15 (83.33) |
| ROS1, No (%)  |                                   | NaN | NaN |
| Mutation      | | 0 | 0 | 0 | 0 | 0 | 0 |
| Wild          | | 35 (27.34) | 7 (16.28) | 28 (32.94) | 33 (67.35) | 20 (64.52) | 13 (72.22) |
| Smoke, No (%) |                                   | 0.78 | 0.74 |
| Never         | | 49 (38.28) | 18 (41.86) | 31 (36.47) | 14 (28.57) | 10 (32.26) | 4 (22.22) |
| Former        | | 79 (61.72) | 25 (58.14) | 54 (63.53) | 35 (71.43) | 21 (67.74) | 14 (77.78) |
| ECOG scale, No (%) | | 0.43 | 0.49 |
| 0             | | 29 (22.66) | 7 (16.28) | 22 (25.88) | 10 (16.33) | 5 (16.13) | 5 (27.78) |
| 1             | | 91 (71.09) | 34 (79.07) | 57 (67.06) | 38 (81.63) | 25 (80.65) | 13 (42.22) |
| ≥2            | | 8 (6.25) | 2 (4.65) | 6 (7.06) | 1 (2.04) | 1 (323) | 0 (0) |
| SUVmax        | | 0.014* | 0.15 |

Continued
Table 2  Continued

| Characteristic | Retrospective MCC ICI-treated patients (N=128) | Prospective MCC ICI-treated patients (N=49) |
|----------------|-----------------------------------------------|---------------------------------------------|
|                | Deeply learned score | P values | Deeply learned score | P values |
| Mean±SD        | All 11.82±6.98 | High (N=43) 13.44±5.83 | Low (N=85) 11.00±7.32 | All 14.59±9.53 | High (N=31) 14.49±6.37 | Low (N=18) 14.77±13.13 |
| Clinical benefit, No (%) | <0.001* | 0.005* |
| DCB            | 69 (53.91) | 34 (79.07) | 35 (65.31) | 7 (38.89) |
| NDB            | 59 (46.09) | 9 (20.93) | 50 (58.82) | 17 (34.69) |
| Progression-free survival | <0.001* | 0.015* |
| Median (95% CI) | 7.43 (6.39 to 8.47) | 15.80 (9.49 to 22.11) | 5.50 (2.87 to 8.13) | 10.50 (6.36 to 14.64) |
| Median (95% CI) | 21.77 (13.50 to 30.03) | 27.60 (NR) | 19.77 (13.82 to 25.72) | 17.00 (NR) |
| Deep learning score | <0.001* | <0.001* |
| Median (IQR) | 0.48 (0.01–0.93) | 0.63 (0.60–0.71) | 0.42 (0.32–0.48) | 0.55 (0.14–0.86) |
|                | (0.56–064) | (0.31–0.45) | (0.31–0.45) | (0.31–0.45) |

*p value<0.05. The comparison of age, BMI, and SUVmax between two groups was performed with Wilcoxon sign rank test, PFS and OS were compared with log-rank test, and the rest variables were compared with Fisher’s exact test.

ADC, adenocarcinoma; ALK, anaplastic lymphoma kinase; BMI, body mass index; DCB, durable clinical benefit; ECOG, Eastern Cooperative Oncology Group Performance Status; EGFR, epidermal growth factor receptor; ICI, immune checkpoint inhibitor; MCC, H Lee Moffitt Cancer Center and Research Institute; NDB, non-durable benefit; NR, not reached; SCC, squamous cell carcinoma; TNM, tumor, node, metastases.
CI: 66.67% to 95.83%), and specificity of 60.00% (95% CI: 20.00% to 100%) (figure 3C and online supplemental table S4).

As another meaningful quantitative index associated with PD-L1 expression validated in other studies,32 SUVmax showed poorer performance to discriminate between PD-L1-positive and PD-L1-negative expression with AUCs of 0.69 (95% CI: 0.62 to 0.75; p<0.001), 0.68 (95% CI: 0.57 to 0.78; p<0.001), 0.66 (95% CI: 0.53 to 0.77; p=0.014), and 0.56 (95% CI: 0.28 to 0.84; p=0.69) in the
Clinical prognostic validation of DLS in ICI treatment

The DLS in the patients experiencing DCB was significantly higher compared with those who did not in both the MCC ICI-treated retrospective (0.54 vs 0.43, p<0.001) and prospective (0.57 vs 0.45, p=0.025) cohorts. The AUCs of the DLS to identify the DCB patients were 0.70 (95% CI: 0.63 to 0.77, p<0.001) and 0.72 (95% CI: 0.62 to 0.84, p=0.014) in the retrospective and prospective...
patients (figure 4A,B), respectively. Similar results were obtained in the external VA test cohort with an AUC of 0.70 (95% CI: 0.52 to 0.88, p=0.040) (figure 4C).

For the retrospective patients, the PFS and OS were significantly longer among patients with high DLS (≥0.55) versus patients with low DLS (PFS: HR=0.41 (95% CI: 0.25 to 0.67, p=0.001); OS: HR=0.48 (95% CI: 0.25 to 0.91, p=0.024); figure 4A). Among patients with high DLS, the median PFS and OS were 15.80 months and 27.60 months compared with 5.37 months and 19.77 months for patients with low DLS (PFS: p<0.001; OS: p=0.021). Similar results were also observed in the prospective patients with high to low DLS ratio-based HRs of 0.38 (95% CI: 0.18 to 0.85, p=0.019) and 0.13 (95% CI: 0.033 to 0.49, p=0.003) for PFS and OS, respectively (figure 4B). High DLS patients had a longer median PFS of 14.33 months compared with 5.00 months in the low DLS patients (p<0.001). Notably, a median time to an OS event was not reached in the high DLS group and was 11.23 months in the low DLS group (p<0.001). The external VA test patients further validated the prognostic value of DLS with HRs of 0.35 (95% CI: 0.12 to 0.99, p=0.047) and 0.23 (95% CI: 0.07 to 0.72, p=0.020) for PFS (9.30 vs 2.37 months, p=0.038) and OS (15.53 vs 4.93 months, p=0.007), respectively (figure 4C).

Multivariable analysis for clinical outcomes prediction

Univariable logistic and Cox regression analyses of the clinical characteristics (online supplemental tables S5–S8) and gene mutation showed that none of three gene mutations were associated with clinical outcome and that

---

**Figure 4** Performance of the DLS in prognosis prediction. (A) The ROC curve of DLS in DCB prediction, and the PFS and OS relative to the DLS (DLS cutoff: 0.55) in the retrospective MCC ICI-treated patients. (B) The ROC curve of DLS in DCB prediction, and the PFS and OS relative to the DLS (DLS cutoff: 0.55) in the prospective MCC ICI-treated patients. (C) The ROC curve of DLS in DCB prediction, and the PFS and OS relative to the DLS (DLS cutoff: 0.55) in the external VA test patients. P value was from log rank test. AUC, area under the receiver operating characteristic curve; DCB, durable clinical benefit; DLS, deeply learned score; HDLS, high DLS; ICI, immune checkpoint inhibitor; LDLS, low DLS; MCC, H Lee Moffitt Cancer Center and Research Institute; OS, overall survival; PFS, progression-free survival; ROC, receiver operating characteristic; VA, James A Haley Veterans' Administration.
patients with lower ECOG status and ADC showed significantly longer OS and PFS. Stratified analyses by histology and ECOG performance status were thus performed to investigate the ability of DLS to predict outcomes in these subgroups. Among patients with ADC, the DCB rates were 91.3% and 100% in patients with high DLS versus 50.88% and 62.5% in patients with low DLS in both retrospective and prospective cohorts (p<0.001), respectively (figure 5). Among SCC patients, though the DCB rates were lower compared with ADC, the patients with higher DLS still had a significantly higher DCB rates in both retrospective and prospective cohorts. Congruously, the PFS and OS of high DLS group were also longer than the low DLS group in both ADC and SCC subgroups (online supplemental table S9). The results of the stratified analysis based on ECOG status (online supplemental table S10) also showed that low DLS was still associated with poor outcomes among patients with high ECOG status (≥1). The above results demonstrated the added value of DLS to the clinical prognostic markers in more accurate quantitative prognosis prediction.

Multivariable logistic regression and Cox proportional hazards regression analyses were conducted to adjust for potential confounding variables. Models including DLS, histology, and ECOG status were developed using the MCC ICI-treated retrospective cohort, which demonstrated high performance statistics (online supplemental table S11 and figure S8) with C-indices of 0.87 (95% CI: 0.83 to 0.92, p<0.001) and 0.82 (95% CI: 0.72 to 0.92, p<0.001) in DCB prediction, 0.73 (95% CI: 0.68 to 0.78, p<0.001) and 0.74 (95% CI: 0.67 to 0.87, p<0.001) in PFS prediction, 0.77 (95% CI: 0.71 to 0.84, p<0.001) and 0.70 (95% CI: 0.50 to 0.87, p<0.001) in OS prediction for the MCC ICI-treated retrospective cohort and independent MCC ICI-treated prospective cohort, respectively, showing better performance than clinical characteristics only models (including ECOG and histology) (p<0.05, online supplemental table S12). These models also

Figure 5 Stratification analysis of the performance of the DLS in prognosis prediction. (A) The DCB rates of the different subgroups of the MCC retrospective and prospective ICI-treated patients. (B) The PFS relative to the DLS and histology in the MCC retrospective and prospective ICI-treated patients. (C) The OS relative to the DLS and histology in the MCC retrospective and prospective ICI-treated patients. Note: HADC is short for HDLS ADC, meaning ADC patients with high DLS; LADC is short for LDLS ADC, meaning ADC patients with low DLS; HSCC is short for HDLS SCC, meaning SCC patients with high DLS; and LSCC is short for LDLS SCC, meaning SCC patients with low DLS, the high DLS versus low DLS defined by 0.55. P value was from log rank test. ADC, adenocarcinoma; DCB, durable clinical benefit; DLS, deeply learned score; ICI, immune checkpoint inhibitor; MCC, H Lee Moffitt Cancer Center and Research Institute; OS, overall survival; PFS, progression-free survival; SCC, squamous cell carcinoma.
demonstrated high performance statistics with C-indices of 0.81 (95% CI: 0.70 to 0.93, p<0.001), 0.70 (95% CI: 0.59 to 0.80, p<0.001), and 0.70 (95% CI: 0.59 to 0.81, p<0.001) in DCB, PFS, and OS prediction, respectively, in the independent external VA patients. The calibration curves of the different models on MCC ICI-treated retrospective cohort, independent MCC ICI-treated prospective cohort, and the external VA cohort provided in online supplemental figure S9 also showed good agreements between the predictions and actual observation.

**DISCUSSION**

PD-L1 expression status based on IHC is currently used as a clinical decision-making tool to support the use of checkpoint inhibitors in patients with NSCLC. Because this relies on invasive biopsies, an alternative non-invasive method to predict PD-L1 status would be useful. In this study, we developed a deep learning model using standard-of-care PET/CT images to measure PD-L1 status non-invasively and showed that the DLS could discriminate between positive and negative expression with an AUC of 0.89 in the SPH training cohort, 0.84 in the SPH validation cohort, and 0.82 in the two independent MCC PD-L1 and VA PD-L1 test cohorts. When the DLS was combined with clinical covariates and tested for clinical utility by identifying patients most likely to benefit from immunotherapy, we found high C-indices of 0.81–0.87 for predicting DCB, but somewhat attenuated C-indices of 0.70–0.77 for the DLS to predict PFS and OS in the MCC ICI-treated retrospective and two independent MCC ICI-treated prospective and VA ICI-treated cohorts.

Others have investigated the utility of radiomics as a non-invasive approach to predict PD-L1 expression status. The current work is a significant advance over these prior studies, which were limited to a single institution, did not validate against independent cohorts, and used cohorts with many early-stage cancers included that are not candidates for ICI therapy. Further, our statistical power outperformed both of these prior studies, which generated AUCs of 0.86 and 0.73, respectively. As an alternative to PD-L1 as a companion biomarker, it should be recognized that tumor mutational burden (TMB), defined as the number of mutations per DNA megabase, is also promising biomarker for predicting immunotherapy responses in patients with advanced stage lung cancer. He et al developed and tested a non-invasive CT-based TMB predictor with 327 patients, which yielded high prognostic value in PFS and OS prediction of immunotherapy in patients with advanced NSCLC. Despite these findings, TMB is not yet a clinically approved diagnostic biomarker attributed in part to the lack of harmonization in panel-based TMB quantification, adequate methods to convert TMB estimates across different panels, and robust predictive cutoff points.

We and others have developed radiomic models to predict lung cancer immunotherapy treatment response regardless of PD-L1 status. Some of these have higher accuracies than do the current model in predicting DCB, PFS, or OS following ICI therapy. However, it is important to note that there is a distinction between the development and application of a completely new companion biomarker such as these, compared with one that provides an alternative assay to assess a currently approved companion biomarker as in the current study. We contend that a radiomic biomarker that predicts PD-L1 status will be more readily accepted into clinical practice compared with a radiomic biomarker that bypasses this known pathway. The current work is the first to develop a PD-L1 radiomic signature and then to use this for response prediction. Additionally, these prior studies were mostly limited to CT and required explicit tumor segmentation, which can render the results to be operator dependent. By contrast, our study utilized deep learning, which did not require accurate tumor segmentation or hard-coded feature extraction and was conducted using rigorous training and validation in multiple cohorts from three institutions. The current study is the single largest multi-institutional radiomic study population of patients with NSCLC to date treated with immunotherapy to predict PD-L1 status and subsequent treatment response using 18F-FDG PET/CT.

In radiomics, it is critical to relate the findings to an underlying biology. One of the high-response areas of the middle layer of the SResCNN model recognized the necrotic region (activation_8_filter_8 in figure 2A,B) through the visualization, suggesting that some final discriminant deeply learned features originate from necrotic regions, which is consistent with the correlation between necrosis and DLS and Jteige’s results. This could be explained with the presence of hypoxia, which can lead to necrotic cell death and upregulate PD-L1 via hypoxia-inducible factor-1α. Additionally, peritumoral regions were also highlighted as informative (activation_8_filter_8 and positive/negative filter in figure 2C,D), which is supported by prior work that higher levels of PD-L1+ staining in cells of peritumoral areas. These findings revealed an advantage of deeply learned models, which can agnostically capture features from the tumor and peritumoral microenvironments. One possible reason for the significant better predictive ability of the hyper-image compared with PET or CT alone may be these two important regions could be better and easier localized by utilizing both metabolic and anatomical information as reflected by PET and CT images, respectively.

We do acknowledge some limitations of this study. First, the PD-L1 prediction training data were limited to a single institution and EGFR mutations were highly prevalent in the Asian patient population at 40% compared with only 7% in whites. This concern is somewhat mitigated by the observation of no significant association between ethnicities and PD-L1 status, and the insignificant different AUCs between mutated EGFR subgroups and wild-type EGFR subgroups. Second, compared with other PD-L1 level detection methods, such as ELISA.
immunofluorescence,47 and flow cytometry,48 only IHC was used in this study to detect PD-L1 expression levels based on the recommendation in the NCCN Clinical Practice Guidelines,5 its ease of use, strong repeatability, and high accuracy.49,50 Comparison among different detection methods should be considered in future research although these other methods are also dependent on biopsy. Third, the patient cohorts were heterogeneous in terms of PET/CT image acquisition. However, this can be viewed as a strength, as this heterogeneity decreases the possibility of overfitting to a specific subset of tumors or imaging parameters, and thus will result in a model that is more robust and transportable. Fourth, the stage distribution was different between the SPH and the MCC cohorts, as the MCC cohort contained more advanced stage patients. To investigate this, we measure the DLS among the subset of SPH patients with advanced stage and obtained high AUCs of 0.90 (95% CI: 0.85 to 0.97, p<0.001), suggesting that stage does not dramatically affect the final DLS prediction.

CONCLUSION
In conclusion, an effective and stable deeply learned score to measure PD-L1 expression status non-invasively was identified and may serve as a prognostic biomarker to guide immunotherapy. Because images are routinely obtained and are not subject to sampling bias per se, we propose that the individualized risk assessment information provided by these analyses may be useful as a future clinical decision support tool pending in larger prospective trials.

Author affiliations
1Department of Cancer Physiology, Moffitt Cancer Center, Tampa, Florida, USA
2Department of Nuclear Medicine, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China
3Department of Radiology, Shengjing Hospital of China Medical University, Shenyang, China
4Department of Thoracic Oncology, Moffitt Cancer Center, Tampa, Florida, USA
5Department of Radiation Oncology, James A. Haley Veterans Affairs Medical Center, Tampa, Florida, USA
6Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Engineering Medicine, Beihang University, Beijing, China
7Institute of Automation, Chinese Academy of Sciences, Beijing, China
8Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, Florida, USA

Contributors
The authors meet criteria for authorship as recommended by the International Committee of Medical Journal Editors. WM, MBS, and RJG contributed to the conception and design of the work; LJ, JT, JEG, and EK contributed to the image and clinical data; LJ, LS, and EK interpreted the images; WM wrote the convolution neural network algorithm and did the statistical analyses; JT, MBS, and RJG supervised the study; MBS and RJG revised the work critical for important intellectual content. All authors contributed to the production of the final manuscript.

Funding
This study was funded by US Public Health Service research grant U01 CA143002 and R01 CA190105 (principal investigator RJG). This material includes work supported with resources and the use of facilities at the James A Haley Veterans’ Hospital.

Competing interests
RJG declared a potential conflict with HealthMyne, Inc (Investor [major], Board of Advisors [uncompensated]). EK has no direct conflict of interest but does have stock ownership in Abbvie, Alexion Pharmaceuticals, Biogen and research clinical trial funding with Advantagene. JEG reports receiving commercial research grants from AstraZeneca, Merck, Array, Epic Sciences, Genentech, Bristol-Myers Squibb, Bi, Trovagene, and Novartis and is a consultant/ advisory board member for AstraZeneca, Janssen, Genentech, Eli Lilly, Celgene, and Takeda, and other remuneration from Genentech, AstraZeneca, Merck, and Lilly/ Genentech.

Patient consent for publication
Not required.

Ethics approval
All procedures performed in studies involving human participants were in accordance with the ethical standards of the Institutional Review Board at the Shanghai Pulmonary Hospital (SPH), University of South Florida (USF), and James A Haley Veterans’ Administration (VA) with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Provenance and peer review
Not commissioned; externally peer reviewed.

Data availability statement
Data are available upon reasonable request. The PET/CT imaging data and clinical information are not publicly available for patient privacy purposes, but are available from the corresponding authors upon reasonable request (RJG and MBS). The remaining data are available within the Article, Supplementary Information or available from the authors upon request. The models and the code used to test and evaluate the model are available on Zenodo (https://doi.org/10.5281/zenodo.4731166).

Supplemental material
This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and it is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access
This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See https://creativecommons.org/licenses/by/4.0/.

ORCID iDs
Wei Mu http://orcid.org/0000-0001-7970-8666
Lei Jiang http://orcid.org/0000-0002-9479-132X
Matthew B Schabath http://orcid.org/0000-0003-3241-3216

REFERENCES
1 Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (oak): a phase 3, open-label, multicentre randomised controlled trial. Lancet 2017;389:255–65.
2 Gandhi L, Rodriguez-Abreu D, Gadgeel S, et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. N Engl J Med 2018;378:2078–92.
3 National Comprehensive Cancer Network (NCCN). NCCN Clinical Practice Guidelines in Oncology. Non-small Cell Lung Cancer version 4.2021, 2021.
4 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;366:2443–54.
5 Yu H, Boyle TA, Zhou C, et al. PD-L1 expression in lung cancer. J Thorac Oncol 2016;11:964–75.
6 Rossi S, Toschi L, Castello A, et al. Clinical characteristics of patient selection and imaging predictors of outcome in solid tumors treated with checkpoint-inhibitors. Eur J Nucl Med Mol Imaging 2017;44:2310–25.
7 Liu Y, Kim J, Qu F, et al. Ct features associated with epidermal growth factor receptor mutation status in patients with lung adenocarcinoma. Radiology 2016;280:271–80.
8 Grossmann P, Stringfeld O, El-Hachem N, et al. Defining the biological basis of radiomic phenotypes in lung cancer. Elife 2017;6. doi:10.7554/eLife.23421. [Epub ahead of print: 21 07 2017].
9 Herbst RS, Soria J-C, Kovarneitz M, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature 2014;515:563–7.
Mu W, et al. J Immunother Cancer 2021;9:e002118. doi:10.1136/jitc-2020-002118

10 Chen DS, Mellman I. Elements of cancer immunity and the cancer-immune set point. Nature 2017;541:321–30.
11 Jreige M, Letovanez I, Chaba K, et al. 18F-FDG PET metabolic-to-morphological volume ratio predicts PD-L1 tumour expression and response to PD-1 blockade in non-small-cell lung cancer. Eur J Nucl Med Mol Imaging 2019;46:1–10.
12 Koh YW, Han J-H, Park SY, et al. Glut1 as a prognostic factor for classical Hodgkin’s lymphoma: correlation with PD-L1 and PD-L2 expression. J Pathol Transl Med 2017;51:152–8.
13 Cui Y, Li Y, Li X. The impact of PD-L1 on glucose metabolism of lung adenocarcinoma cells. J Nucl Med 2018;59:1252–52.
14 Damgaci S, Ibrahim-Haisham A, Enríquez-Navas PM, et al. Hypoxia and acidosis: immune suppressors and therapeutic targets. Immunology 2018;154:354–62.
15 Lopci E, Toschi L, Grizi F, et al. Correlation of metabolic information on FDG-PET with tissue expression of immune markers in patients with non-small cell lung cancer (NSCLC) who are candidates for upfront surgery. Eur J Nucl Med Mol Imaging 2016;43:1954–61.
16 Wu X, Huang Y, Zhao Q, et al. PD-L1 expression correlation with metabolic parameters of FDG-PET/CT and clinicopathological characteristics in non-small cell lung cancer. EJNMMI Res 2020;10:1–9.
17 Mu W, Jiang L, Zhang J, et al. Non-invasive decision support for NSCLC treatment using PET/CT radiomics. Nat Commun 2020;11:5228.
18 Health UDo, Services H. Software as a medical device (SAMD): clinical evaluation. Guidance for industry and Food and Drug Administration Staff, 2017.
19 Rizvi NA, Hellmann MD, Snyder A, et al. Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 2015;348:124–8.
20 Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer 2009;45:228–47.
21 Rizvi H, Sanchez-Vega F, La K, et al. Molecular Determinants of Response to Anti-Programmed Cell Death (PD)-1 and Anti-Programmed Death-Ligand 1 (PD-L1) Blockade in Patients With Non-Small-Cell Lung Cancer Profiled With Targeted Next-Generation Sequencing. J Clin Oncol 2018;36:633–41.
22 Takada K, Okarnoto T, Toyokawa G, et al. The expression of PD-L1 protein as a prognostic factor in lung squamous cell carcinoma. Lung Cancer 2017;104:7–15.
23 Schroeder W, Ng L, Cates J. The ITK software guide, 2003.
24 Bican J, Janeba D, Táborská K, et al. Image overlay using alpha-blending technique. Nucl Med Rev Cent East Eur 2002;5:53.
25 Chollet F. Deep learning with python and Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek. MITP-Verlags GmbH & Co. KG, 2018.
26 Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on computer vision 2017.
27 DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837.
28 Rovere MTL, Bigger JT, Marcus FL, et al. Baroreflex sensitivity and heart-rate variability in prediction of total cardiac mortality after myocardial infarction. The Lancet 1998;351:478–84.
29 Diedenhofen B, Musch J. cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 2015;10:e0121945.
30 Gainor JF, Shaw AT, Sequist LV, et al. EGFR mutations and ALK rearrangements are associated with low response rates to PD-1 pathway blockade in non-small cell lung cancer: a retrospective analysis. Clin Cancer Res 2016;22:4585–93.
31 Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017;14:749–62.
32 Chen R, Zhou X, Liu J, et al. Relationship between the expression of PD-L1/PD-1 and 18F-FDG uptake in bladder cancer. Eur J Nucl Med Mol Imaging 2019;46:848–54.
33 Jiang M, Sun D, Guo Y, et al. Assessing PD-L1 expression level by radiomic features from PET/CT in non-small cell lung cancer patients: an initial result. Acad Radiol 2020;27:171–9.
34 Rizvi NA, Hellmann MD, Snyder A, et al. Classic Driver Oncogene Mutations in EGFR, ALK, or ROS1 and Correlation of radiomic features with PD-L1 expression in early stage non-small cell lung cancer (ES-NSSCLC) to predict recurrence and overall survival (OS). J Clin Oncol 2018;36:e24247–e47.
35 Yu Y, Zeng D, Ou Q, et al. Association of survival and immune-related biomarkers with immunotherapy in patients with non-small cell lung cancer: a meta-analysis and individual patient-level analysis. JAMA Netw Open 2019;2:e1906879–e79.
36 He B, Dong D, She Y, Di Dong YS, Zhou C, et al. Predicting response to immunotherapy in advanced non-small-cell lung cancer using tumor mutational burden radiomic biomarker. J Immunother Cancer 2020;8:600050.
37 Fancello L, Gandini S, Pellicci PG, et al. Tumor mutational burden quantification from targeted gene panels: major advancements and challenges. J Immunother Cancer 2019;7:183.
38 Tunali I, Gray JE, Qi J, et al. Novel clinical and radiomic predictors of rapid disease progression phenotypes among lung cancer patients treated with immunotherapy: an early report. Lung Cancer 2019;129:75–9.
39 Tunali I, Tan Y, Gray JE. Hypoxia-Related radiomics predict immunotherapy response: a multi-cohort study of NSCLC. bioRxiv.
40 Trebeschi S, Drago SG, Birkbak NJ, et al. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol 2019;30:998–1004.
41 Sun R, Limkin EJ, Valyamaou M, et al. A radiomics approach to assess tumour-infiltrating CD8+ cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol 2018;19:1180–91.
42 Semenza GL. Hypoxia-Inducible factors in physiology and medicine. Cell 2012;149:999–1009.
43 Chang C-H, Qiu J, O’Sullivan D, et al. Metabolic competition in the tumor microenvironment is a driver of cancer progression. Cell 2015;162:1229–41.
44 Sun S, Fei X, Mao Y, et al. PD-1+ immune cell infiltration inversely correlates with survival of operable breast cancer patients. Cancer Immunol Immunother 2014;63:395–406.
45 Rangachari D, VanderLaan PA, Shea M, et al. Correlation between Classic Driver Oncogene Mutations in EGFR, ALK, or ROS1 and 22C3-PD-L1 ≥50% Expression in Lung Adenocarcinoma. J Thorac Oncol 2017;12:878–83.
46 Enninga EAL, Harrington SM, Creedon DJ, et al. Immune checkpoint molecules soluble program death ligand 1 and galectin-9 are increased in pregnancy. Am J Reprod Immunol 2018;79:e12795.
47 Velchetti V, Schalper KA, Carvajal DE, et al. Programmed death-ligand 1 expression in non-small cell lung cancer. Lab Invest 2014;94:107–16.
48 Caldwell C, Johnson CE, Balaji VN, et al. Identification and validation of a PD-L1 binding peptide for determination of PDL1 expression in tumors. Sci Rep 2017;7:1–11.
49 Hirsch FR, McElhinny A, Stanforth D, et al. PD-L1 immunohistochemistry assays for lung cancer: results from phase 1 of the blueprint PD-L1 IHC assay comparison project. J Thorac Oncol 2017;12:208–22.
50 Tsao MS, Kerr KM, Koczk M, et al. PD-L1 immunohistochemistry comparability study in real-life clinical samples: results of blueprint phase 2 project. J Thorac Oncol 2018;13:1302–11.
Supplementary data

Supplementary Methods
Supplemental S1: Inclusion criteria for SPH and MCC cohorts

For PD-L1 prediction, patients with informed consent were accrued from the Shanghai Pulmonary Hospital (SPH), Shanghai, China, between January 2017 and June 2018 with following inclusion criteria were included: 1) histologically confirmed primary NSCLC; 2) pathological examination of PD-L1 status before any treatment; 3) PET/CT scans obtained within one month before biopsy (Bx) for immunohistochemistry (IHC) and no treatments was performed during this interval; 5) baseline clinical characteristics (including age, sex, stage, histology, and smoking history) and gene (EGFR, ALK and ROS1) mutation status were available. Based on these inclusion criteria, 400 patients were identified and subsequently assigned to a training cohort (SPH-training, N = 284) and an independent test cohort (SPH-test, N = 116). Using the same inclusion criteria, 85 NSCLC patients were accrued from H. Lee Moffitt Cancer Center & Research Institute (MCC), Tampa, FL and were used as external independent test cohort for prediction of PD-L1 expression (MCC-PD-L1 cohort).

For the distinct cohorts to predict patient response and outcomes, 128 patients were identified with histologically confirmed advanced stage (stage IIIB and IV) NSCLC who were treated with immunotherapy (anti-PD-L1 or anti-PD-1) between June 2011 and December 2017 at MCC using the following criteria: 1) PET/CT images were available during the interval (less than 6 months) of the last treatment (or diagnosis) and the start of immunotherapy; 2) no other treatment were performed during the interval; 3) follow-up time was greater than 6 months; and 4) no immune-related severe adverse events (Grade according to Common Terminology Criteria for Adverse Events (CTCAE)≥3$^5$) were observed or reported during treatment; 5) baseline clinical characteristics (including age, sex, stage, histology, Eastern Cooperative Oncology Group (ECOG) scale, brain metastasis status, and smoking history) and gene (EGFR, ALK and ROS1) mutation status were available. Using the same inclusion criteria, a prospective validation cohort was curated of 49 NSCLC patients who were treated with immunotherapy between January 2018 to June 2019.

For the external VA cohort to validate the DLS and the prognostic models, 35 patients with available PET/CT images from 72 patients with advanced stage NSCLC treated with immunotherapy (anti-PD-L1 or anti-PD-1) between July 2015 and February 2019 were identified according to the above criteria.

The progression of the distinct ICI-treated cohorts used to investigate the association of the DLS and clinical characteristics with the clinical outcome including DCB (PFS≥6month$^{20}$), PFS, and OS, was defined using Response Evaluation Criteria in Solid Tumors (RECIST1.1)$^{21}$. For PFS, an event was defined as death or either...
clinical or RECIST based progression of cancer and the data were right-censored at 6 years and 1.5 years for the retrospective and prospective cohorts, respectively. For OS, an event was defined as death and the data were right censored at 6 years and 1.5 years for the retrospective and prospective cohorts, respectively. Because we don’t have as much follow-up time for the prospective cohort, these two cohorts have different censoring values. The index date for both OS and PFS was the date of initiation of immunotherapy.

[S1]. Institute, N. C. Common terminology criteria for adverse events (CTCAE) v4.0, 2010.

Supplemental S2: PD-L1 expression by immunohistochemistry (IHC)

To ensure a reliable PD-L1 IHC score, all patients in this study underwent surgical resection or biopsy of the primary tumor using standardized protocol. To reduce the sampling artifact, the portion of the tumor specimen was carefully examined, and the portion with more malignant cells, less differentiated cells, and less hemorrhage was subjected to histopathological confirmation within 2 weeks after the $^{18}$F-FDG PET/CT scan. Furthermore, routine IHC analysis was performed to determine PD-L1 expression in all the lesions of SPH cohort and MCC-PD-L1 cohort using the same antibody. For the SPH cohort, the platform of Dako Link 48 and the antibody of Dako 22C3 (Agilent Cat# GE00621-2, RRID:AB_2833074) were used for PD-L1 staining to quantify the presence of PD-L1. For the MCC-PD-L1 cohort, the PD-L1 22C3 mouse monoclonal antibody (Agilent Cat# GE00621-2, RRID:AB_2833074) purchased from Dako, was performed utilizing the Dako EnVision FLEX visualization system on the Dako Autostainer Link 48. To compensate for reader bias, all the staining results were reviewed and analyzed by 2 experienced pathologists who were blinded to each other’s scores and unaware of the patients’ clinical information. When there was discrepancy, the two pathologists would have a mutual discussion to reach a consensus.

Supplemental S3: Details of the training of the deep learning model

Details of the small-residual-convolutional-network

The SResCNN is similar to the Resnet50 but with fewer layers and smaller number of residual blocks, and the architecture was shown in supplemental Figure S1. The architecture was comprised with one convblock (including a $3 \times 3$ convolutional layer followed by a batch normalization layer and a rectified linear unit (ReLU) activation layer), 8 residual blocks (Resblock), and one fully connected layer. Finally, a softmax activation layer was connected to the last fully connected layer, which was used to yield the prediction probabilities of nodule...
candidates. To prevent overfitting, one dropout layer with probability of 0.5 was added to the fully connected layers. Additionally, the model was optimized using the binary cross entropy loss function.

**Preparation of the input images**

After registration using ITK-SNAP, a square or an irregular box, which was close to the boundary of the tumor, was delineated manually in ITK software firstly by experienced nuclear medicine radiologist, and then the input regions of interest (ROIs) could be generated automatically after, dilation, resize and fusion to keep the entire tumor and its peripheral region were included (Supplemental Figure S3). For each slice of the tumor, the area of the smallest square mask (SSM) including the delineated region was regarded as the area of the tumor in this slice (Supplemental Figure S3B). Because of the big difference of the central slice and peripheral slices, only the slices with the area larger than the 30% of the maximum tumor cross-sectional area of this patient were regarded as valid input images and were used as the input of the deep learning model. The area here means the area of the smallest square including the delineated region (Supplemental Figure S3C). 10,650 ROIs were generated for training. In order to keep the training data had more balanced label, cubic spline interpolation of the adjacent two slices from the same patient with positive PD-L1 expression was used to generate new augmented slices on the condition that each patient was used with the same times, and finally 14,011 ROIs (6,722 PD-L1 positive and 7,289 PD-L1 negative) were used as the training dataset.

All ROIs were resized to the same size (64×64) using cubic spline interpolation and were standardized by z-score normalization before input to the model. As such, the input ROI was subtracted from the mean intensity value and divided by the standard deviation of the image intensity, before inputting to the deep learning model, to reduce the offset effect due to different equipment and different reconstruction parameters. 

**Training of the SResCNN network**

The training of the model focuses on the optimization of the parameters of the SResCNN model to build a relationship between PET/CT images and PD-L1 expression status (positive: 1 or negative: 0). We employed binary cross entropy as the loss function and the Adam optimizer with an initial learning rate = 0.0001, beta_1=0.9, beta_2=0.999. The learning rate was reduced by a factor of 5 if no improvement of the loss of the validation dataset was seen for a ‘patience’ number (n=50) of epochs. The batch size was set to 64. During the training, augmentation including width/height-shift, horizontal/vertical-flip, rotation and zoom were used to reduce overfitting. The training was stopped after waiting an additional 50 epochs since the validation loss stopped to degrade. The learning rate and batch size was determined with five-fold-cross validation under the patient level, and the combination that yielded the best average accuracy on the validation folds was chosen.
Application of the SResCNN network

The generated ROI-based hyper-image was input into the SResCNN model after z-score normalization, and a deeply learned score (DLS) representing the PD-L1 positivity could be yielded after a sequential activation of convolution and pooling layers. To develop a robust prediction, all valid slices of each patient were fed into the SResCNN model and the average DLSs with equal weight for each slice was regarded as the final PD-L1 positive probability of the tumor.

[S2]. Oh, Shu Lih, et al. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Computers in biology and medicine 102 (2018): 278-287.

[S3] COONEY, Ciaran; FOLLI, Raffaella; COYLE, Damien. Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC). IEEE, 2019: 1311-1316.

Supplemental S4: Correlation investigation between necrosis and DLS

First, necrosis in PET images was defined as an area of hypometabolism within the hypermetabolic tumor (i.e., classically a rim of hypermetabolism with a hypometabolic center)\textsuperscript{S4}. Hypometabolism (necrosis) is defined as the region with SUV less than 42% of the maximum SUV\textsuperscript{39}. Then, the ratio of necrosis to global lesion volume of the PET images (termed the necrosis-to-global volume ratio, NVR) was calculated and expressed as percentage to quantify the necrosis.

Thus, the relation between necrosis and the DLS could be investigated by calculating the Spearman's correlation, and linear regression between NVR and DLS, and only the cases with necrosis regions were included for this experiment.

[S4] Rakheja R, Makis W, Tulbah R, Skamene S, Holcroft C, Nahal A, et al. Necrosis on FDG PET/CT correlates with prognosis and mortality in sarcomas. Am J Roentgenol 2013;201:170-7

Supplemental S5: Radiomic quality score (RQS)

Radiomics is a rapidly maturing field in machine learning. To rigorously assess the quality of study design, Lambin et al. developed a 36-point “Radiomics Quality Score” (RQS) metric that evaluates 16 different key components\textsuperscript{31}. The full list of criteria is described in Supplemental Table S2, which shows that the current study had a RQS of 22. To put this in perspective, a recent meta-analysis\textsuperscript{S5} analyzed 77 radiomics publications
and documented that the mean ± S.D. RQS across all studies was 9.4 ± 5.6, indicating that the current study is in the upper 5 percentage of radiomics study designs.

[55]. Park JE, Kim D, Kim HS et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 2019.
**Supplementary Figures**

**Figure S1. Study design and inclusion and exclusion diagram.** The SPH data, which included PD-L1 expression data and the corresponding imaging data, was used to develop a deeply learned score (DLS) as a surrogate IHC based PD-L1 status. The MCC-PD-L1 data which included PD-L1 expression data, the corresponding imaging data and the treatment information was used for the clinical association and analytic validation analyses of the DLS through the measurement accuracy and the prognostic comparison, while the larger MCC retrospective and prospective cohorts comprised of patients treated with ICI were used for further clinical validation. Then the MCC ICI-treated retrospective and prospective cohorts were further used for developing and testing the prognosis prediction models that included other prognosis-related clinical variables. The external VA cohort was used as a further validation for PD-L1 status measurement and prognosis evaluation to ICI treatment.
Figure S2. Illustration of the ResCNN model. This model is composed of convolutional layers with kernel size 3x3, batch normalization, pooling, and drop out layers. Note. /2 means the convolution layer of the Convblock with stride of 2.

Figure S3. Illustration of the generation of the input hyper-image. A square or an irregular box, which was close to the boundary of the tumor, was delineated manually in ITK software firstly, and then the hyper-image was generated after dilation, resize and fusion automatically.
Figure S4. Two different digital simulated phantoms were constructed as A and B. To show the importance of the fusion images, A and B were kept to have the same heterogeneity distribution. Entropy and Inverse Difference calculated from 3D co-occurrence matrix were used to measure the heterogeneity and the homogeneity of the phantoms. From A1 and B1 (simulated PET images), A2 and B2 (simulated CT images), the two phantoms have the same heterogeneity and homogeneity distribution. But from A3 and B3, the two phantoms could be identified based on different heterogeneity and homogeneity, which means the fusion images could reflect the relative different positional relationship of the heterogeneity. Using this image to construct a 3-channel hyper-image together with PET and CT images was more convenient for the training of the ResCNN model in one hand. In the other hand, incorporating the prior knowledge into the model can decrease the size of the deep learning model and limit the risk of overfitting.
**Figure S5. Distribution of the DLS and SUVmax across the PD-L1 expression.** The p-values in the down (up)-right corner of each plot are from the Spearman’s correlation analysis and ANOVA analysis. The p-values in the bridge between different cohorts are from the pairwise post hoc tests.
Figure S6. Correlation of the DLS and necrosis-to-global volume ratio within the patients possessing necrotic regions.

Figure S7. Distribution of the DLS across the patient cohorts. The bootstrapped mean value of DLSs as well as the standard error bar for different cohorts.
Figure S8. Nomograms for multivariable regression. (A) DCB nomogram obtained with multivariable logistic regression for DCB prediction (e.g., for a ADC patient with DLS of 0.6 and ECOG 1, his total points are 78 (DLS 0.6 corresponding to point 0, ECOG 1 corresponding to point 50, ADC corresponding to point 28, 0+50+28=78), which corresponds to a DCB probability of 0.40); (B) PFS nomogram obtained with multivariable Cox proportional hazards regression for PFS prediction (e.g., for a ADC patient with DLS of 0.6 and ECOG 1, his total points are 50 (DLS 0.6 corresponding to point 0, ECOG 1 corresponding to point 50, ADC corresponding to point 0, 0+50+0=50), which corresponds to a 6-month PFS probability of 0.85, 1-year PFS probability of 0.68, and 2-year PFS probability of 0.59). (C) OS nomograms obtained with multivariable Cox proportional hazards regression for OS prediction (e.g., for a SCC patient with DLS of 0.6 and ECOG 1, his total points are 74 (DLS 0.6 corresponding to point 0, ECOG 1 corresponding to point 50, SCC corresponding to point 24, 0+50+24=74), which corresponds to a 6-month OS probability of 0.87, 1-year OS probability of 0.63, and 2-year OS probability of 0.37).
Figure S9. Calibration Plots for the multi-variable models. (A) The assessment of the DCB prediction model calibration in the MCC ICI-treated retrospective cohort (intercept = 0, slope = 0.99, C-index = 0.87), MCC ICI-treated prospective cohort (intercept = 0.067, slope = 0.73, C-index = 0.82), and the external VA ICI-treated cohort (intercept =-0.42, slope =0.61, C-index =0.81). (B) The assessment of the PFS prediction model calibration in the MCC ICI-treated retrospective cohort (6-month: slope = 1.00 [95%CI: 0.69-1.18], 1-year: slope = 1.00 [95%CI: 0.76-1.16], 2-year: slope = 1.00 [95%CI: 0.73-1.23], C-index = 0.73), MCC ICI-treated prospective cohort (6-month: slope = 1.00 (95%CI: 0.28-1.45), 1-year: slope = 1.00 (95%CI: 0.09-1.37), C-index=0.74), and the external VA ICI-treated cohort (6-month: slope = 1.00 (95%CI: 0.40-1.79), 1-year: slope = 1.00 (95%CI: 0.07-1.78) , C-index=0.70). (C) The assessment of the OS prediction model calibration in the MCC ICI-treated retrospective cohort (6-month: slope = 1.00 [95%CI: 0.56-1.32], 1-year: slope = 1.00 [95%CI: 0.58-1.21],2-year: slope = 1.00 [95%CI: 0.67-1.21], C-index=0.74 ), MCC ICI-treated prospective cohort (6-month: slope = 1.00 [95%CI: -0.099-1.44], 1-year: slope = 1.00 [95%CI: 0.12-1.39] , C-index = 0.70), and the external VA ICI-treated cohort (6-month: slope = 1.00 [95%CI: 0.28-1.36], 1-year: slope = 1.00 [95%CI: 0.19-1.30] , C-index = 0.70).
### Supplementary Tables

**Table S1. Acquisition parameters for the PET/CT imaging for each cohort**

| Characteristic                        | SPH MCC-PD-L1 cohort | MCC retrospective ICI-treated cohort | MCC prospective ICI-treated cohort | VA cohort |
|--------------------------------------|----------------------|-------------------------------------|----------------------------------|-----------|
| **Manufacturer, No. (%)**            |                      |                                     |                                  |           |
| SIEMENS/CPS                          | 400 (100)            | 64 (75.29)                          | 19 (14.84)                       | 12 (24.49)| 35(100)   |
| GE Medical                           | 0                    | 17 (20.00)                          | 103 (80.47)                      | 37 (75.51)| 0         |
| PHILIPS                              | 0                    | 4 (4.71)                            | 6 (4.69)                         | 0         | 0         |
| **Kilovoltage peak (kVp), No. (%)**  |                      |                                     |                                  |           |
| 120                                  | 400 (100)            | 77 (90.59)                          | 118 (92.19)                      | 44 (89.80)| 35(100)   |
| 130                                  | 0                    | 5 (5.88)                            | 7 (5.47)                         | 3 (6.12)  | 0         |
| 140                                  | 0                    | 3 (3.53)                            | 3 (2.34)                         | 2 (4.08)  | 0         |
| **Reconstruction method, No. (%)**   |                      |                                     |                                  |           |
| OSEM                                 | 0                    | 22 (25.88)                          | 38 (21.69)                       | 12 (24.49)| 35(100)   |
| PSF+TOF                              | 400 (100)            | 4 (4.71)                            | 7 (5.47)                         | 2 (4.08)  | 0         |
| VPHD                                 | 0                    | 15 (17.65)                          | 6 (4.69)                         | 16 (32.65)| 0         |
| 3D IR                                | 0                    | 43 (50.59)                          | 73 (57.03)                       | 19 (38.78)| 0         |
| 'BLOB-OS-TF'                         | 0                    | 1 (1.18)                            | 4 (3.13)                         | 0         | 0         |
| **Current (mA)**                     |                      |                                     |                                  |           |
| Median (range)                       | 193 (90-463)         | 83 (31-238)                         | 85 (27-299)                      | 85 (29-299)| 97(53-134) |
| **Interval between administration and image acquisition** | | | | | |
| Mean ± SD                            | 62.68±12.13          | 95.39±18.84                         | 96.26±24.03                      | 96.06 ± 22.81| 93.15±18.34 |
| **Dosage MBq/kg**                    |                      |                                     |                                  |           |
| Mean ± SD                            | 3.70 ± 0.32          | 6.07 ± 2.11                         | 6.03 ± 1.87                      | 5.97±1.87 | 6.27±1.25 |
| **PET Slice Thickness**              |                      |                                     |                                  |           |
| Median (range)                       | 5                    | 3.27(3.26-5)                        | 3.27(3.27-5)                     | 3.27(3.26-5)| 5         |
| **PET Pixel Spacing**                |                      |                                     |                                  |           |
| Median (range)                       | 4.07                 | 5.31(2.74-4.67)                     | 5.47(2.73-4.67)                  | 4.07(3.65-4.67)| 4.07     |
| **CT Slice Thickness**               |                      |                                     |                                  |           |
| Median (range)                       | 3                    | 3.27(3.26-5)                        | 3.375(3.27-5)                    | 3.27(3.27-5)| 2         |
| **CT Pixel Spacing**                 |                      |                                     |                                  |           |
| Median (range)                       | 0.9766               | 1.37(0.88-1.37)                     | 1.37(0.88-1.37)                  | 1.37(0.98-1.37)| 0.9766   |

a. The PET/CT scanners of PHILIPS include GEMINI TF TOF16 and GEMINI TF TOF 16. The PET/CT scanners of SIEMENS include Biograph 6, Biograph 40, Biograph 64 and Emotion Duo. The PET/CT scanners of GE Medical include Discovery 600, Discovery STE and Discovery ST. The PET/CT scanners of CPS is 1080.
| Criteria                                                                 | Points system                                      | Maximal score | Actual score of this work |
|-------------------------------------------------------------------------|----------------------------------------------------|---------------|---------------------------|
| Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability | + 1 (if protocols are well-documented) + 1 (if public protocol is used) | 2             | 1                         |
| Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities | 1                                                  |               | 0                         |
| Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability | 1                                                  |               | 0                         |
| Imaging at multiple time points - collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage) | 1                                                  |               | 0                         |
| Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features | – 3 (if neither measure is implemented) + 3 (if either measure is implemented) | 3             | 0                         |
| Multivariable analysis with non radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating /inferencing between radiomics and non radiomics features | 1                                                  |               | 1                         |

Table S2. The criteria and maximal radiomic quality score as well as the actual score of this work
| Method                                                                 | Points |
|-----------------------------------------------------------------------|--------|
| Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology | 1      |
| Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results | 1      |
| Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation) | + 1 (if a discrimination statistic and its statistical significance are reported) + 1 (if a resampling method technique is also applied) |
| Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation) | + 1 (if a calibration statistic and its statistical significance are reported) + 1 (if a resampling method technique is also applied) |
| Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker | + 7 (for prospective validation of a radiomics signature in an appropriate trial) |
| Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance | - 5 (if validation is missing) + 2 (if validation is based on a dataset from the same institute) + 3 (if validation is based on a dataset from another institute) + 4 (if validation is based on two datasets from two distinct institutes) + 4 (if the study validates a previously published signature) + 5 (if validation is based on three or more datasets from distinct institutes) |
| Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics | 2 | 2 | 2 |
| Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis) | 2 | 2 | 0 |
| Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated) | 1 | 1 | 0 |
| Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study + 1 (if scans are open source) +1 (if region of interest segmentations are open source) +1 (if code is open source) + 1 (if radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source) | 4 | 3 |
| Total score | 36 | 22 |
Table S3. Demographic and clinical characteristics of VA ICI-treated patients

| Characteristic                      | All (N=35) | Deep Learning Score | P  |
|-------------------------------------|------------|---------------------|----|
|                                     |            | High (N=28) | Low (N=7) |    |
| Age (y)                             |            | 71.40±7.19 | 71.77±7.36 | 71.29±9.05 | 0.76 |
| Mean ± SD                           |            |            |            |            |
| Sex, NO. (%)                        | NaN        |            |            |            |
| Male                                | 35 (100)   | 26 (100)  | 7 (100)   |    |
| Female                              | 0          | 0         | 0         |    |
| TNM stage                           |            | 10 (28.57) | 9 (31.03) | 1 (16.67)  | 0.64 |
| III                                 | 25 (71.43) | 19 (68.97) | 6 (83.33) |    |
| IV                                  |            |            |            |            |
| Histology (baseline), NO. (%)       | 0.19       |            |            |            |
| ADC                                 | 19 (54.29) | 14 (50.00) | 5 (83.33) |    |
| SCC                                 | 16 (45.71) | 14 (51.72) | 2 (16.67) |    |
| EGFR, NO. (%)                       | NaN        |            |            |            |
| Mutation                            | 0          | 0         | 0         |    |
| Wild                                | 22 (62.86) | 17 (58.62) | 5 (83.33) |    |
| ALK, NO. (%)                        | 0.74       |            |            |            |
| Mutation                            | 0          | 0         | 0         |    |
| Wild                                | 22 (62.86) | 17 (58.62) | 5 (83.33) |    |
| Smoke, NO. (%)                      | 1.00       |            |            |            |
| Never                               | 1 (2.86)   | 1 (3.45)  | 0         |    |
| Former                              | 34 (97.14) | 27 (96.55) | 9 (100)   |    |
| ECOG Scale, NO. (%)                 | 0.42       |            |            |            |
| 0                                   | 7 (20.00)  | 5 (17.24) | 2 (33.33) |    |
| 1                                   | 22 (62.86) | 18 (62.07) | 4 (66.67) |    |
| >=2                                 | 6 (17.14)  | 5 (20.69) | 1         |    |
| Clinical Benefit, NO. (%)           | 0.032      |            |            |            |
| DCB                                 | 19 (54.29) | 18 (62.07) | 1 (16.67) |    |
| NDB                                 | 16 (45.71) | 10 (37.93) | 6 (83.33) |    |
| Progression-free Survival           |            |            |            |            |
| Median (95%CI)                      | 8.13(4.50,11.77) | 9.30 (4.69,13.91) | 2.37 (1.60,3.13) | 0.038* |
| Overall Survival                    |            |            |            |            |
| median (95%CI)                      | 13.10 (5.67,20.53) | 15.53 (9.54,21.53) | 4.93 (2.20,7.67) | 0.007* |
| PD-L1 by IHC                        |            |            |            |            |
| Positive                            | 24 (68.57) | 20 (72.41) | 4 (50.00) | 0.007* |
| Negative                            | 5 (14.29)  | 2 (6.90)  | 3 (50.00) |    |
| Deep Learning Score                 |            |            |            |            |
| Median (IQR)                        | 0.63 (0.56,0.66) | 0.64 (0.58,0.67) | 0.37 (0.28,0.46) | <.001* |

Note. PD-L1 expression status was significantly associated with DCB (p=0.017, Fisher’s Exact Test).
### Table S4. Predictive performance of the ResCNN model by histology subtypes

|                | DLS                                      | ADC | SCC | ADC + SCC | Internal (VA PD-L1) | External (VA PD-L1) |
|----------------|-----------------------------------------|-----|-----|-----------|---------------------|--------------------|
| **AUC [95%CI]**| SPH-training cohort                      |     |     |           | MCC-PD-L1-test cohort |                   |
| DLS ADC       | 0.89 [0.84,0.94]                         | 0.81[0.71,0.91] | 0.88[0.76,0.96] | 0.89 [0.64,1.00]    |                     |
| DLS SCC       | 0.87 [0.80,0.95]                         | 0.89 [0.76,1.00] | 0.77 [0.61,0.91] | 0.80 [0.50,1.00]    |                     |
| DLS ADC + SCC | 0.89 [0.85,0.93]                         | 0.84 [0.76,0.92] | 0.82 [0.74,0.89] | 0.82 [0.65,0.98]    |                     |
| **SUVmax**    | SPH-training cohort                      |     |     |           | MCC-PD-L1-test cohort |                   |
| ADC           | 0.66 [0.57,0.74]                         | 0.72 [0.59,0.84] | 0.63 [0.44,0.79] | 0.68 [0.43,0.93]    |                     |
| SCC           | 0.67 [0.54,0.79]                         | 0.41 [0.20,0.61] | 0.71 [0.53,0.86] | 0.47 [0.09,0.99]    |                     |
| ADC + SCC     | 0.69 [0.62,0.75]                         | 0.68 [0.57,0.78] | 0.66 [0.53,0.77] | 0.56 [0.28,0.84]    |                     |
| **Accuracy**  | SPH-training cohort                      |     |     |           | MCC-PD-L1-test cohort |                   |
| DLS ADC       | 83.33 [78.20,88.23]                      | 79.76 [71.43,87.47] | 82.98 [70.34,93.62] | 81.25 [62.50,100]   |                     |
| DLS SCC       | 77.50 [67.50,86.25]                      | 75.00 [59.38,87.50] | 71.05 [55.26,84.21] | 76.92 [53.85,92.31] |                     |
| DLS ADC + SCC | 81.69 [77.11,85.91]                      | 78.45 [71.55,85.83] | 77.65 [69.41,85.88] | 79.31 [65.52,93.10] |                     |
| **SUVmax**    | SPH-training cohort                      |     |     |           | MCC-PD-L1-test cohort |                   |
| ADC           | 75 [69.12,79.41]                         | 60.71 [50.00,71.43] | 59.57 [44.68,72.34] | 68.75 [37.50,81.25] |                     |
| SCC           | 60 [50,71.25]                            | 43.75 [28.12,62.5] | 57.89 [42.11,71.05] | 53.85 [30.77,76.92] |                     |
| ADC + SCC     | 70.77 [64.79,75.35]                      | 67.24 [59.48,75]  | 62.35 [47.06,68.24] | 55.83 [34.48,68.97] |                     |
| **Sensitivity**| SPH-training cohort                      |     |     |           | MCC-PD-L1-test cohort |                   |
| DLS ADC       | 83.33 [70.83,93.75]                      | 63.16 [42.11,84.21] | 73.08 [55.87,88.46] | 78.57 [57.14,100]   |                     |
| DLS SCC       | 86.49 [75.68,94.59]                      | 81.25 [56.25,100] | 63.64 [40.91,81.82] | 90 [70,100]         |                     |
| DLS ADC + SCC | 84.71 [76.47,91.76]                      | 77.43 [57.14,85.71] | 68.75 [55.26,81.25] | 83.33 [66.67,95.83] |                     |
| **SUVmax**    | SPH-training cohort                      |     |     |           | MCC-PD-L1-test cohort |                   |
| ADC           | 72.92 [31.25,95.83]                      | 68.42 [47.37,89.47] | 69.23 [50.84,62.6] | 64.29 [28.57,78.57] |                     |
| SCC           | 67.57 [35.14,78.38]                      | 37.50 [12.50,62.5] | 50.00 [31.82,68.18] | 50.00 [20.00,80.00] |                     |
| ADC + SCC     | 52.94 [35.29,85.88]                      | 42.86 [28.57,60.00] | 60.42 [47.06,68.23] | 45.83 [29.17,62.5]  |                     |

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s).

J Immunother Cancer

doi: 10.1136/jitc-2020-002118

Mu W, et al. J Immunother Cancer 2021; 9:e002118. doi: 10.1136/jitc-2020-002118
Note. Cutoff for DLS is 0.55 for ADC cohort, SCC cohort and ADC+SCC cohort for all three cohorts. Cutoffs for SUV are 6.8, 14.5 and 12.11 for ADC cohort, SCC cohort and ADC+SCC cohort for all three cohorts, respectively, according to the ROC curves of training cohort.

Table S5. Univariable analysis of risk factors for DCB prediction

|                | Retrospective | Prospective |
|----------------|---------------|-------------|
|                | Odds Ratio (95% CI) | p        | Odds Ratio (95% CI) | p         |
| Age            | 1.02 (0.99-1.05) | 0.11       | 0.97 (0.92-1.04) | 0.39      |
| BMI            | 1.13 (1.04-1.22) | 0.002      | 0.87 (0.76-1.00) | 0.051     |
| Sex            | 0.86 (0.43-1.70) | 0.66       | 1.62 (0.49-5.32) | 0.43      |
| Stage          | 0.74 (0.30-1.79) | 0.50       | -                   | -         |
| Brain Metastasis | 0.20 (0.02-1.86) | 0.16       | 1.50 (0.42-5.32) | 0.53      |
| Histology(baseline) | 0.39 (0.19-0.82) | 0.013     | 0.06 (0.013-0.27) | <.001     |
| EGFR           | 0.74 (0.17-3.14) | 0.68       | -                   | -         |
| ALK            | 0.75 (0.045-12.30) | 0.84     | -                   | -         |
| ROS1           | -               | -          | -                   | -         |
| Smoking status | 0.54 (0.26-1.12) | 0.096      | 1.64 (0.46-5.87) | 0.45      |
| ECOG           | 0.055(0.012-0.24) | <.001     | 0.77 (0.17-3.44) | 0.73      |
| SUVmax         | 1.01 (0.96-1.06) | 0.74       | 1.11 (1.00-1.23) | .047      |

Note., For Sex: male was assigned 1 and female was assigned 2; for histology, ADC was assigned 1 and SCC was assigned 2.

Table S6. Univariable analysis of risk factors for PFS

|                | Retrospective | Prospective |
|----------------|---------------|-------------|
|                | Hazard ratio (95% CI) | p        | Hazard ratio (95% CI) | p         |
| Age            | 0.99(0.98-1.01) | 0.53       | 1.03(0.99-1.07) | 0.21      |
| BMI            | 0.96 (0.92-1.00) | 0.048      | 1.05 (0.96-1.15) | 0.25      |
| Sex            | 0.97 (0.65-1.44) | 0.86       | 0.76 (0.36-1.62) | 0.48      |
| Stage          | 1.04 (0.62-1.72) | 0.89       | 1.99 (0.59-6.71) | 0.27      |
| Brain Metastasis | 1.55 (0.5-4.25) | 0.40       | 0.62 (0.22-1.8) | 0.38      |
| Histology(baseline) | 2.13 (1.40-3.24) | <.001     | 6.22 (2.70-14.308) | <.001     |
| EGFR           | 0.49 (0.15-1.56) | 0.23       | 0.043 (0.41-73) | 0.37      |
| ALK            | 1.31 (0.32-5.36) | 0.71       | -                   | -         |
| ROS1           | 0.53 (0.07-4.05) | 0.54       | -                   | -         |
| Smoking status | 1.04 (0.67-1.60) | 0.88       | 0.95 (0.42-2.15) | 0.89      |
| ECOG           | 2.40 (1.38-4.19) | 0.002      | 0.82 (0.35-1.95) | 0.66      |
| SUVmax         | 1.00 (0.97-1.03) | 0.97       | 0.99 (0.95-1.03) | 0.62      |

Note., For Sex: male was assigned 1 and female was assigned 2; for histology, ADC was assigned 1 and SCC was assigned 2.
| Risk Factor                  | Retrospective Hazard ratio (95% CI) | p   | Prospective Hazard ratio (95% CI) | p   |
|-----------------------------|-------------------------------------|-----|-----------------------------------|-----|
| Age                         | 0.99 (0.97-1.01)                    | 0.31| 1.01 (0.96-1.07)                  | 0.74|
| BMI                         | 0.93 (0.88-0.99)                    | 0.017| 1.01 (0.91-1.13)                  | 0.80|
| Sex                         | 0.87 (0.51-1.49)                    | 0.87| 0.86 (0.29-2.57)                  | 0.79|
| Stage                       | 1.24 (0.60-2.56)                    | 0.56| 1.29 (0.27-6.09)                  | 0.75|
| Brain Metastasis            | 2.17 (0.67-7.01)                    | 0.20| 1.26 (0.34-4.63)                  | 0.73|
| Histology (baseline)        | 2.44 (1.38-4.30)                    | 0.002| 5.34 (1.40-20.32)                 | 0.014|
| EGFR                        | 1.09 (0.26-4.59)                    | 0.91| 0.043 (0-1950)                    | 0.57|
| ALK                         | 0.045 (0-68.25)                     | 0.41| -                                 | -   |
| ROS1                        | 21.10 (0-Inf)                       | 0.77| -                                 | -   |
| Smoking status              | 1.64 (0.72-2.50)                    | 0.35| 1.11 (0.34-3.64)                  | 0.87|
| ECOG                        | 6.24 (1.94-20.11)                   | 0.002| 0.75 (0.23-2.47)                  | 0.63|
| SUVmax                      | 1.01 (0.97-1.04)                    | 0.74| 0.97 (0.90-1.05)                  | 0.48|

Note. For Sex: male was assigned 1 and female was assigned 2; for histology, ADC was assigned 1 and SCC was assigned 2.
Table S8. Clinical characteristics associated with patient outcomes

| Characteristic | K-M Analysis MCC retrospective ICI cohort | Cox regression MCC prospective ICI cohort | K-M Analysis MCC prospective ICI cohort | Cox regression MCC retrospective ICI cohort |
|---------------|------------------------------------------|----------------------------------------|-----------------------------------------|------------------------------------------|
|               | median time (IQR), months                | p                                      | median time (IQR), months              | p                                        |
|               |                                          |                                        |                                        |                                           |
| Histology     |                                          |                                        |                                        |                                           |
| ADC SCC       | 10.60 [3.73, 50.20] 3.93 [1.63, 8.40]   | <.001*                                  | 17.00 [7.93, NR] 4.00 [3.00, 5.73]     | <.001*                                   |
|               |                                          |                                        |                                        |                                           |
| OS ADC SCC    | NR [12.13, NR] 11.07 [6.47, 27.60]       | 0.002*                                  | NR [17.00, NR] 11.43 [11.23, NR]       | .006*                                    |
|               |                                          |                                        |                                        |                                           |
| ECOG PFS      |                                          |                                        |                                        |                                           |
| 0 >=1         | 12.47 [7.67, NR] 5.50 [0.3, 15.80]       | .001*                                  | 8.37 [5.76, 17.00] 7.93 [3.93, -]      | .065                                     |
|               |                                          |                                        |                                        |                                           |
| OS 0 >=1      | NR [23.87, NR] 15.10 [6.57, NR]          | <.001*                                  | 17.00 [11.23, NR] NR [11.43, NR]       | .63                                      |
|               |                                          |                                        |                                        |                                           |
| Note: NR means the median (or 25%, or 75%) of survival has been not yet reached.
| Histology | MCC retrospective IO cohort | MCC prospective IO cohort |
|-----------|----------------------------|--------------------------|
| ADC       | ADC                        | SCC                      |
| High DLS  | Low DLS                    | High DLS                 |
| (N=23)    | (N=57)                     | (N=20)                   |
|          | High DLS                   | Low DLS                  |
| ADC       | Low DLS                    | ADC                      |
| High DLS  | (N=20)                     | Low DLS                  |
| (N=28)    | High DLS                   | Low DLS                  |
| (N=20)    | High DLS                   | Low DLS                  |
| (N=8)     | High DLS                   | Low DLS                  |
| (N=11)    | High DLS                   | Low DLS                  |
| (N=10)    |                            |                          |
| PFS       |                            |                          |
| HR [95%CI]| 0.26 [0.12, 0.58]          | 0.42 [0.22, 0.81]        |
|           | 0.30 [0.093, 0.97]         | 0.54 [0.18, 1.61]        |
| p (cox)   | 0.001*                     | 0.010*                   |
| p (K-M)   | <0.001*                    | 0.008*                   |
| OS        |                            |                          |
| HR [95%CI]| 0.38 [0.13, 1.12]          | 0.48 [0.21, 1.06]        |
|           | 0.087 [0.008, 0.90]        | 0.29 [0.058, 1.44]       |
| p (cox)   | 0.080                      | 0.068                    |
| p (K-M)   | 0.069                      | 0.061*                   |
| Durable Clinical benefit | | |
| DCB rate  | 91.30%                     | 50.88%                   |
|           | 65.00%                     | 21.43%                   |
|           | 100.00%                    | 62.50%                   |
|           | 45.45%                     | 45.45%                   |
|           | 20.00%                     |                          |
| ORR[95%CI]| 10.14 [2.17, 47.32]        | 6.81 [1.88, 24.69]       |
|           | -                          | 3.33 [0.47, 23.47]       |
| p(logistic)| 0.003*                    | 0.004*                   |

Note. : * means P value <.05
### Table S10. DLS and clinical outcomes stratified by ECOG performance status

| ECOG     | MCC retrospective IO cohort | MCC prospective IO cohort |
|----------|-----------------------------|---------------------------|
|          | ECOG=0                      | ECOG>=1                   | ECOG=0                  | ECOG>=1                  |
|          | High DLS (N=7)              | Low DLS (N=22)            | High DLS (N=36)         | Low DLS (N=63)           | High DLS (N=5)           | Low DLS (N=5)           | High DLS (N=26)         | Low DLS (N=13)          |
| PFS      |                             |                           |                          |                           |                           |                           |                           |                           |
| HR [95%CI]| 0.15 [0.019, 1.14]         | 0.38 [0.23, 0.64]         | 0.37 [0.11, 2.73]        | 0.22 [0.09, 0.55]        |
| p (cox)  | 0.066                       | <.001*                    | 0.25                    | 0.001*                   |
| p (K-M)  | 0.025                       | <0.001*                   | 0.21                    | <.001*                   |
| OS       |                             |                           |                          |                           |                           |                           |                           |                           |
| HR[95%CI]| 0.019 [0.00, 0.229.61]     | 0.45 [0.24, 0.87]         | 0.78 [0.069, 8.88]      | 0.048 [0.006, 0.40]     |
| p (cox)  | 0.41                        | 0.009*                    | 0.84                    | 0.005*                   |
| p (K-M)  | 0.12                        | 0.014*                    | 0.83                    | <.001*                   |
| Durable Clinical benefit |         |                           |                          |                           |                           |                           |                           |
| DCB rate | 100%                        | 90.91%                    | 74.29%                  | 23.81%                   | 80.00%                   | 60.00%                   | 80.77%                   | 30.77%                   |
| ORR[95%CI]| -                          | 9.60 [3.71, 24.86]        | 2.67 [0.16, 45.14]      | 9.45 [2.05, 43.61]       |
| p(logistic)| -                          | <.001                     | 0.50                    | 0.004                    |

Note., * means P value <.05
Table S11. Multivariable logistic regression and Cox regression analyses

|          | DCB prediction | PFS estimation | OS estimation |
|----------|----------------|----------------|---------------|
|          | B   | Odds Ratio (95% CI) | p  | B   | Hazard Ratio (95% CI) | p  | B   | Hazard Ratio (95% CI) | p  |
| DLS      | 3.00 | 20.13 (5.71-71.00) | <.001 | -1.37 | 0.25 (0.15-0.42) | .002 | -1.01 | 0.36 (0.19-0.69) | 0.002 |
| Histology| -1.91 | 0.15 (0.045-0.49) | <.001 | 1.12  | 3.06 (1.95-4.80) | <.001 | 1.03  | 2.79 (1.58-4.95) | <.001 |
| ECOG     | -3.41 | 0.033 (0.008-0.14) | <.001 | 1.26  | 3.52 (2.17-5.69) | <.001 | 2.13  | 8.37 (3.91-17.92) | <.001 |
| Constant | 4.54 | <.001           |      |      |                |      |      |                |      |

C-Index (95%CI, p-value)

|          | Retrospective | Prospective | VA  |
|----------|---------------|-------------|-----|
|          | 0.87 (0.82-0.92, <.001) | 0.84 (0.74-0.94, <.001) | 0.81 (0.70-0.93, <.001) |
|          | 0.73 (0.68-0.78, <.001) | 0.74 (0.67-0.87, <.001) | 0.70 (0.59-0.80, <.001) |
|          | 0.77 (0.71-0.84, <.001) | 0.70 (0.50-0.87, 0.02) | 0.70 (0.59-0.81, <0.001) |

AIC

|          | Retrospective | Prospective | VA |
|----------|---------------|-------------|----|
|          | 118.12        | 49.58       | 40.50 |
|          | 703.23        | 173.36      | 130.16 |
|          | 371.19        | 83.57       | 105.75 |
### Table S12. Multivariable logistic regression and Cox regression analysis only using clinical characteristics

|               | DCB prediction | PFS estimation | OS estimation |
|---------------|----------------|----------------|---------------|
|               | B              | Odds Ratio     | p             | B             | Hazard Ratio | p  | B             | Hazard Ratio | p  |
| ECOG          | -2.54          | 0.079 (0.023-0.27) | <.001         | 0.98          | 2.68(1.65-4.34) | <.001 | 2.11          | 8.11 (3.73-17.65) | <.001 |
| Histology     | -0.96          | 0.38 (0.17-0.87)   | .023          | 0.79          | 2.20(1.44-3.37) | <.001 | 0.95          | 2.58 (1.46-4.56) | 0.001 |
| Constant      | 3.71           | 0.38 (0.17-0.87)   | <.001         | 0.79          | 2.20(1.44-3.37) | <.001 | 0.95          | 2.58 (1.46-4.56) | 0.001 |
| C-Index (95% CI, p-value) |               |                 |               |               |               |         |               |               |     |
| Retrospective | 0.75 (0.69-0.81, <.001) | 0.67 (0.62-0.72, <.001) | 0.74 (0.67-0.81, <.001) | | | | | |
| Prospective   | 0.72 (0.60-0.85, 0.004) | 0.61 (0.50-0.72, 0.049) | 0.60 (0.44-0.76, 0.24) | | | | | |
| VA            | 0.70 (0.57-0.84, 0.014) | 0.61 (0.50-0.72, 0.052) | 0.67 (0.54-0.79, 0.008) | | | | | |
| AIC           |                |                 |               |               |               |         |               |               |     |
| Retrospective | 149.73         | 733.37          | 379.75        | | | | | |
| Prospective   | 62.50          | 188.03          | 87.57         | | | | | |
| VA            | 46.65          | 134.27          | 109.04        | | | | | |
| P value of Z-test compared with the models in Table S11 |               |                 |               |               |               |         |               |               |     |
| Retrospective | <.001          | 0.024           | 0.001         | | | | | |
| Prospective   | 0.050          | 0.009           | 0.001         | | | | | |
| VA            | 0.029          | 0.076           | 0.036         | | | | | |