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

Wei Mu, Lei Jiang, Yu Shi, Ilke Tunali, Jhanelle E Gray, Evangelia Katsoulakis, Jie Tian, Robert J Gillies, Matthew B Schabath

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) of biopsies is the only clinically approved companion biomarker to trigger the use of ICI therapy. 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 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. 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) 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). However, measuring PD-L1 by IHC requires surgical or biopsied tumor specimens, which are collected through invasive procedures and associated with risk of morbidities. 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.

Radiomic analyses of quantitative image features based on shape, size, voxel intensity, and texture are strongly associated with gene and protein expression in NSCLC. Signatures are typically extracted from the intratumoral region, but it is becoming increasingly appreciated that the peritumoral region, 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, which is significantly correlated with metabolic rate, GLUT-1 expression, pAKT levels, hypoxia, and acidosis. 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.\textsuperscript{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.\textsuperscript{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 $^{18}$F-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),\textsuperscript{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).

**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 DLS 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 months)20, PFS, and OS, were defined using Response Evaluation Criteria in Solid Tumors (RECIST V.1.1).20 The index date for both OS and PFS was the date of initiation of immunotherapy.

The requirement for informed consent was waived by the IRBs, as no personal health information is reported.

18F-FDG PET/CT imaging
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.

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-SNAP23 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 SSMSs 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,24 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.26 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 method27 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.28 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 cocor.29 Given prior research has suggested that PD-L1 expression is negatively correlated with EGFR mutation status,30 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 calculated31 (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 al17). 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:...
### Table 1  Demographic and clinical characteristics of patients used to measure PD-L1 status

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>SPH training cohort (N=284)</th>
<th>SPH validation cohort (N=116)</th>
<th>MCC PD-L1 test cohort (N=85)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PD-L1+</td>
<td>PD-L1−</td>
<td>P values</td>
</tr>
<tr>
<td>Age (years)</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean±SD</td>
<td>62.71±8.78</td>
<td>63.51±8.56</td>
<td></td>
</tr>
<tr>
<td>Sex, No (%)</td>
<td>0.035*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>58 (68.24)</td>
<td>108 (54.27)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>27 (31.76)</td>
<td>91 (45.73)</td>
<td></td>
</tr>
<tr>
<td>TNM stage</td>
<td>0.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>43 (50.59)</td>
<td>122 (61.31)</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>22 (26.19)</td>
<td>34 (17.08)</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>11 (13.10)</td>
<td>31 (15.58)</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>9 (10.71)</td>
<td>12 (6.03)</td>
<td></td>
</tr>
<tr>
<td>Histology (baseline), No (%)</td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADC</td>
<td>48 (56.47)</td>
<td>156 (78.39)</td>
<td></td>
</tr>
<tr>
<td>SCC</td>
<td>37 (43.53)</td>
<td>43 (21.61)</td>
<td></td>
</tr>
<tr>
<td>EGFR, No (%)</td>
<td>0.020*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutation</td>
<td>24 (28.57)</td>
<td>87 (43.72)</td>
<td></td>
</tr>
<tr>
<td>Wild</td>
<td>55 (65.48)</td>
<td>101 (50.75)</td>
<td></td>
</tr>
<tr>
<td>ALK, No (%)</td>
<td>0.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutation</td>
<td>1 (1.19)</td>
<td>1 (0.50)</td>
<td></td>
</tr>
<tr>
<td>Wild</td>
<td>78 (92.86)</td>
<td>187 (93.97)</td>
<td></td>
</tr>
<tr>
<td>ROS1, No (%)</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mutation</td>
<td>0</td>
<td>2 (1.01)</td>
<td></td>
</tr>
<tr>
<td>Wild</td>
<td>77 (91.67)</td>
<td>179 (89.95)</td>
<td></td>
</tr>
<tr>
<td>Smoking status, No (%)</td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>32 (37.65)</td>
<td>118 (59.3)</td>
<td></td>
</tr>
<tr>
<td>Former</td>
<td>53 (62.35)</td>
<td>81 (40.7)</td>
<td></td>
</tr>
<tr>
<td>SUVmax</td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean±SD</td>
<td>12.32±5.99</td>
<td>8.42±4.83</td>
<td></td>
</tr>
<tr>
<td>Deeply learned score (DLS)</td>
<td>&lt;0.001*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (IQR)</td>
<td>0.70 (0.60–0.78)</td>
<td>0.43 (0.30–0.55)</td>
<td></td>
</tr>
<tr>
<td>PD-L1 positivity by IHC</td>
<td>Continued</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 $r_{pb}$=0.27, p<0.001), male sex (point-biserial $r_{pb}$=0.19, p<0.001), smoking status (point-biserial $r_{pb}$=0.20, p<0.001), and negatively correlated with EGFR status (point-biserial $r_{pb}$=-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 $r_{pb}$=-0.25, p=0.035).

Further, multivariable linear regression (adjusted $r^2=0.15$, 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 $r^2=0.24$, 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: 76.47% 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 83.33% (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) |  |  |
| --- | --- | --- |  | --- | --- | --- |
|  | All | Deeply learned score | All | Deeply learned score | All | Deeply learned score |
|  | High (N=43) | Low (N=85) | P values | High (N=31) | Low (N=18) | P values |
| 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±5.46 |  | 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 (37.50) | 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) | 28 (36.35) |  | 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

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Retrospective MCC ICI-treated patients (N=128)</th>
<th>Prospective MCC ICI-treated patients (N=49)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deeply learned score</td>
<td>Deeply learned score</td>
</tr>
<tr>
<td></td>
<td>All High (N=43)</td>
<td>All High (N=31)</td>
</tr>
<tr>
<td></td>
<td>Mean±SD All</td>
<td>Mean±SD All</td>
</tr>
<tr>
<td></td>
<td>11.82±6.98</td>
<td>14.59±9.53</td>
</tr>
<tr>
<td></td>
<td>Low (N=85)</td>
<td>Low (N=18)</td>
</tr>
<tr>
<td></td>
<td>11.00±7.32</td>
<td>14.49±6.37</td>
</tr>
<tr>
<td></td>
<td>P values</td>
<td>P values</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td>0.005*</td>
</tr>
<tr>
<td>DCB</td>
<td>69 (53.91)</td>
<td>32 (65.31)</td>
</tr>
<tr>
<td></td>
<td>34 (79.07)</td>
<td>25 (80.65)</td>
</tr>
<tr>
<td>NDB</td>
<td>59 (46.09)</td>
<td>17 (34.69)</td>
</tr>
<tr>
<td></td>
<td>9 (20.93)</td>
<td>6 (19.35)</td>
</tr>
<tr>
<td></td>
<td>P values</td>
<td>P values</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td>0.015*</td>
</tr>
<tr>
<td>Median (95% CI)</td>
<td>7.43 (6.39 to 8.47)</td>
<td>5.50 (2.87 to 8.13)</td>
</tr>
<tr>
<td></td>
<td>15.80 (9.49 to 22.11)</td>
<td>10.50 (6.36 to 14.64)</td>
</tr>
<tr>
<td>Overall survival</td>
<td>21.77 (13.50 to 30.03)</td>
<td>19.77 (13.82 to 25.72)</td>
</tr>
<tr>
<td></td>
<td>27.60 (NR)</td>
<td>17.00 (NR)</td>
</tr>
<tr>
<td></td>
<td>Median (95% CI)</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>0.021*</td>
<td>0.015*</td>
</tr>
<tr>
<td>Deep learning score</td>
<td>27.00 (NR)</td>
<td>11.23 (6.69 to 15.78)</td>
</tr>
<tr>
<td></td>
<td>Median (95% CI)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.48 (0.01–0.93)</td>
<td>0.55 (0.14–0.86)</td>
</tr>
<tr>
<td></td>
<td>0.63 (0.60–0.71)</td>
<td>0.59 (0.56–0.64)</td>
</tr>
<tr>
<td></td>
<td>P values</td>
<td>P values</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td>0.001*</td>
</tr>
<tr>
<td></td>
<td>&lt;0.001*</td>
<td></td>
</tr>
</tbody>
</table>

*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: 0.55 to 0.77; p=0.014), and 0.56 (95%CI: 0.28 to 0.84; p=0.69) in the
SPH prediction by PD-L1


Figure 3  Performance of the DLS in predicting PD-L1 status. (A) The distribution of DLS between PD-L1-positive (+) and PD-L1-negative (−) groups in SPH training, SPH validation, and external MCC PD-L1 test cohorts. (B) The receiver operating characteristic curves of DLS and SUV max in SPH training, SPH validation, and external MCC PD-L1 test cohorts. (C) The quantitative performance metrics in SPH training, SPH validation, external MCC PD-L1 test, and external VA test cohorts. ACC, accuracy; AUC, area under receiver operating characteristics curve; DLS, deeply learned score; PD-L1, programmed death-ligand 1; SUV, standardized uptake value; MCC, H Lee Moffitt Cancer Center and Research Institute; SEN, sensitivity; SPEC, specificity; SPH, Shanghai Pulmonary Hospital; VA, James A Haley Veterans’ Administration.

Since histology was found to be significantly associated with PD-L1 status (p<0.01), a stratified analysis was conducted to assess the SResCNN model in predicting PD-L1 status by histology (online supplemental table S4). The results from these analyses indicated this model also performed well in both adenocarcinoma (ADC) and SCC lung cancers. Additionally, although there was significant negative association between DLS and EGFR mutation status, the DLS yielded high AUCs of 0.90 (95% CI: 0.84 to 0.98), 0.87 (95% CI: 0.73 to 1.00), and 0.80 (95% CI: 0.42 to 1.00) in patients with an EGFR mutation and also high, but not statistically significant different AUCs of 0.88 (95% CI: 0.83 to 0.94, DeLong test p=0.71), 0.85 (95% CI: 0.75 to 0.95, DeLong test p=0.61), and 0.82 (95% CI: 0.70 to 0.93, DeLong test p=0.93) in patients with wild type of EGFR in the SPH training, SPH validation, and MCC PD-L1 test cohorts, respectively.

Performance of the PD-L1 DLS was further investigated by comparing its prognostic to that of PD-L1 IHC. Among 85 patients in the MCC PD-L1 test cohort and 29 patients in the external VA test cohort that have PD-L1 IHC data, the DLS achieved AUCs of 0.70 (95% CI: 0.60 to 0.82) and 0.68 (95% CI: 0.47 to 0.89) in DCB prediction for MCC and VA, respectively. These were nearly identical to the DLSs achieved for ADC and SCC in SPH training, SPH validation, and external MCC PD-L1 test cohorts. The DLSs were obtained accordingly. The ICC of these two DLSs was as high as 0.85 (95% CI: 0.80 to 0.90, p<0.001). Further, similar bootstrapped mean values of the DLSs were found across the five patient cohorts (online supplemental figure S7).

Regarding the stability of the DLS, though accurate segmentations were not needed, radiologists had to delineate a rough ROI that contained the tumors and some surrounding tissue. To investigate the effect of the minor differences between the different radiologists in selecting the rough ROIs, the ROIs of the SPH validation patients (n=116 cases) were generated by two radiologists, and two DLSs were obtained accordingly. The ICC of these two DLSs was as high as 0.85 (95% CI: 0.80 to 0.90, p<0.001).

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: 0.03 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
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 Jreige'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 aagnostically 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, only IHC was used in this study to detect PD-L1 expression levels based on the recommendation in the NCCN Clinical Practice Guidelines, its ease of use, strong repeatability, and high accuracy. 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, JY, 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, Troyavage, 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 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**


