Original Article

Artificial intelligence-based analysis for immunohistochemistry staining of immune checkpoints to predict resected non-small cell lung cancer survival and relapse

Haoyue Guo[1,2#], Li Diao[3#], Xiaofeng Zhou[4#], Jie-Neng Chen[5], Yue Zhou[3], Qiyu Fang[1], Yayi He[1], Rafal Dziadziuszko[6], Caicun Zhou[1], Fred R. Hirsch[7]

[1]Department of Medical Oncology, Shanghai Pulmonary Hospital, Tongji University Medical School Cancer Institute, Tongji University School of Medicine, Shanghai, China; [2]School of Medicine, Tongji University, Shanghai, China; [3]Department of Automation, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China; [4]School of Information Management & Engineering, Shanghai University of Finance and Economics, Shanghai, China; [5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA; [6]Department of Oncology and Radiotherapy, Medical University of Gdansk, ul. M. Sklodowskiej-Curie 3A, Gdańsk 80-210, Województwo pomorskie, Poland; [7]Center for Thoracic Oncology, Mount Sinai Cancer, New York, NY, USA

Contributions: (I) Conception and design: H Guo, L Diao, Y Zhou, Q Fang, Y He, C Zhou, FR Hirsch; (II) Administrative support: C Zhou, FR Hirsch; (III) Provision of study materials or patients: R Dziadziuszko, Y He; (IV) Collection and assembly of data: L Diao, X Zhou, Y Zhou, JN Chen; (V) Data analysis and interpretation: L Diao, X Zhou, Y Zhou, JN Chen; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Yayi He, MD, PhD. Department of Medical Oncology, Shanghai Pulmonary Hospital, Tongji University Medical School Cancer Institute, Tongji University School of Medicine, Shanghai, China. Email: 2250601@qq.com.

Background: Conventional analysis of single-plex chromogenic immunohistochemistry (IHC) focused on quantitative but spatial analysis. How immune checkpoints localization related to non-small cell lung cancer (NSCLC) prognosis remained unclear.

Methods: Here, we analyzed ten immune checkpoints on 1,859 tumor microarrays (TMAs) from 121 NSCLC patients and recruited an external cohort of 30 NSCLC patients with 214 whole-slide IHC. EfficientUnet was applied to segment tumor cells (TCs) and tumor-infiltrating lymphocytes (TILs), while ResNet was performed to extract prognostic features from IHC images.

Results: The features of galectin-9, OX40, OX40L, KIR2D, and KIR3D played an un-negatable contribution to overall survival (OS) and relapse-free survival (RFS) in the internal cohort, validated in public databases (GEPIA, HPA, and STRING). The IC-Score and Res-Score were two predictive models established by EfficientUnet and ResNet. Based on the IC-Score, Res-Score, and clinical features, the integrated score presented the highest AUC for OS and RFS, which could achieve 0.9 and 0.85 in the internal testing cohort. The robustness of Res-Score was validated in the external cohort (AUC: 0.80–0.87 for OS, and 0.83–0.94 for RFS). Additionally, the neutrophil-to-lymphocyte ratio (NLR) combined with the PD-1/PD-L1 signature established by EfficientUnet can be a predictor for RFS in the external cohort.

Conclusions: Overall, we established a reliable model to risk-stratify relapse and death in NSCLC with a generalization ability, which provided a convenient approach to spatial analysis of single-plex chromogenic IHC.

Keywords: Tumor microenvironment (TME); tumor-infiltrating lymphocyte (TIL); immune checkpoint; prognosis; deep learning

Submitted Feb 04, 2021. Accepted for publication Apr 16, 2021.
doi: 10.21037/tlcr-21-96
View this article at: http://dx.doi.org/10.21037/tlcr-21-96

Introduction

Non-small-cell lung cancer (NSCLC) remains the top global reason for cancer-relevant deaths (1). Immunotherapy has evolved into the most promising cancer treatment strategies for NSCLC, accompanied by surprising therapeutic results of immune checkpoint inhibitors (ICIs) (2-4). Immune checkpoints pathways such as programmed cell death receptor-1 (PD-1)/programmed cell death ligand-1 (PD-L1), lymphocyte activation gene-3 (LAG-3)/major histocompatibility complex class II (MHC-II), T cell immunoglobulin-3 (TIM-3)/galectin-9, tumor necrosis factor ligand superfamily member 4 (TNFSF4, OX40)/tumor necrosis factor receptor superfamily member 4 (TNFRSF4, OX40L), KIR2D, and KIR-3D from killer cell immunoglobulin-like receptors (KIRs) are modifiers in the immunomodulatory mechanism, which act as a switch for activation of T cells, natural killer cells (NK cells) and other immune cells (4,5). However, tumor cells could escape immune surveillance by unregulated expressing immune checkpoint molecules.

Most current studies exerted quantitative analysis on immune checkpoints, especially PD-1/PD-L1 (6), which is also a significant biomarker approved by the Food and Drug Administration (FDA) for the response of ICIs. Moreover, the existence and distribution of tumor-infiltrating lymphocytes (TILs) are related to a preferable survival and a strengthened efficacy to cancer treatments in multiple cancers, including NSCLC (7,8). These findings prompted the proposal of a series of immunohistochemistry (IHC)-based predictive scores (9-11) in various cancers. However, most of these predictive scores only involve quantitative analysis, but not the spatial location of immune cells or immune checkpoints expression (12-16).

Multiplex immunohistochemistry (mIHC) has been considered a potential tool to reveal cell-cell interactions on a single section (17). Due to the spectral crosstalk in the mIHC, researchers need to purchase expensive hardware and software to visualize and analyze multiple biomarkers one by one (18). What is more, when multiple target proteins co-localize on the same cell, the cross-color interference caused by overlapping signals poses a huge challenge for mIHC (19). Limited by these deficiencies, most institutions have not achieved the conditions for performing the mIHC. However, due to the excessive workload of manually labeling tumor cells (TCs) and TILs in each whole slide, spatial analysis is still challenging to achieve in single-plex chromogenic IHCs. With the advance of medical artificial intelligence (20-23), making full use of single-plex chromogenic IHC techniques for spatial analysis between cells is critical to breaking through these economic and technical limitations.

To better reveal the function of immune checkpoints in the tumor microenvironment (TME), we applied a pattern recognition algorithm that identifies four types of TCs and TILs from single-plex chromogenic IHC sections, based on cell morphology and color. Upon cell segmentation, quantitative and spatial analysis of immune checkpoint expressions were carried out. Meanwhile, we performed another deep learning algorithm to extract prognostic features from IHC images to assist the prediction of immune checkpoint features in survival and relapse (*Figure 1*). We present the following article in accordance with the MDAR reporting checklist (available at http://dx.doi.org/10.21037/tlcr-21-96).

Methods

Ethics statement

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). It was approved by Shanghai Pulmonary Hospital Ethics Committee (approval number: 15-235), and the written informed consent was obtained from all patients.

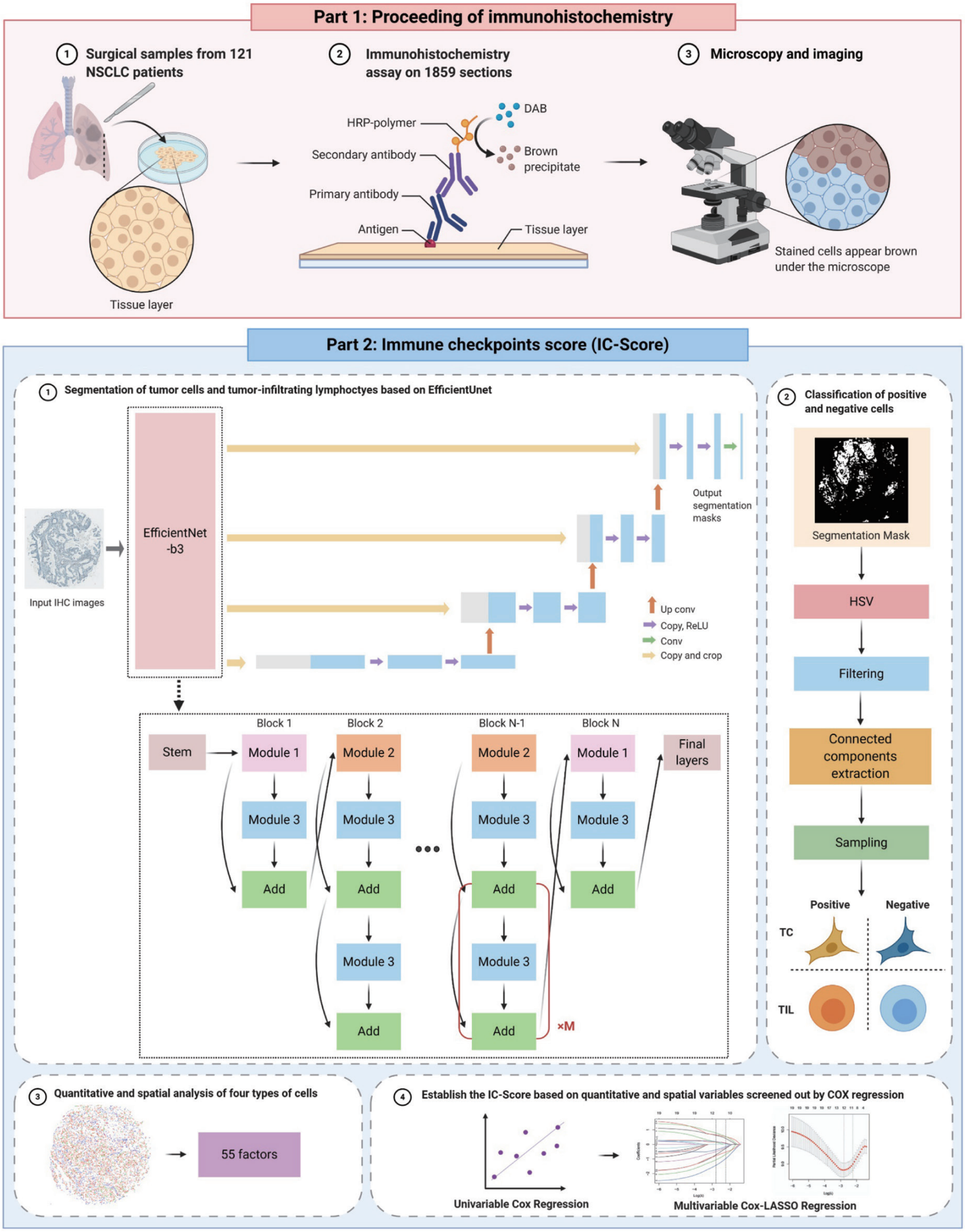
Human subjects

A total of 121 NSCLC patients with 1,859 tissue micro-array (TMAs) images were included as the internal cohort at the Department of Oncology and Radiotherapy, Medical University of Gdansk, Poland, between April 2008 and August 2010. For the external cohort, 30 NSCLC patients with 214 whole-slide images were included at the Department of Medical Oncology, Shanghai Pulmonary Hospital in August 2018. The follow-up deadline for the internal and external cohort was March 2016 and October 2020, respectively.

All patients were diagnosed as resectable NSCLC and had never received any treatment before the surgery. The patients who did not meet the diagnosis or lacked complete follow-up information were excluded. The tumor, nodes, and metastasis (TNM) stage of the internal cohort and external cohort was in term of the 7th and 8th editions of the TNM classification, respectively.

Single-plex chromogenic IHC staining

The single-plex chromogenic IHC was performed as


Part 1: Proceeding of immunohistochemistry
1 Surgical samples from 121 NSCLC patients
Tissue layer
2 Immunohistochemistry assay on 1859 sections
DAB
HRP-polymer
Secondary antibody
Brown precipitate
Primary antibody
Antigen
Tissue layer
3 Microscopy and imaging
Stained cells appear brown under the microscope
Part 2: Immune checkpoints score (IC-Score)
1 Segmentation of tumor cells and tumor-infiltrating lymphoctyes based on EfficientUnet
Input IHC images
EfficientNet -b3
Output segmentation masks
Up conv
Copy, ReLU
Conv
Copy and crop
Block 1
Block 2
Block N-1
Block N
Stem
Module 1
Module 2
Module 2
Module 1
Final layers
Module 3
Module 3
Module 3
Module 3
Add
Add
Add
Add
Module 3
Module 3
Add
Add
×M
2 Classification of positive and negative cells
Segmentation Mask
HSV
Filtering
Connected components extraction
Sampling
Positive
Negative
TC
TIL
3 Quantitative and spatial analysis of four types of cells
55 factors
4 Establish the IC-Score based on quantitative and spatial variables screened out by COX regression
Univariable Cox Regression
Multivariable Cox-LASSO Regression

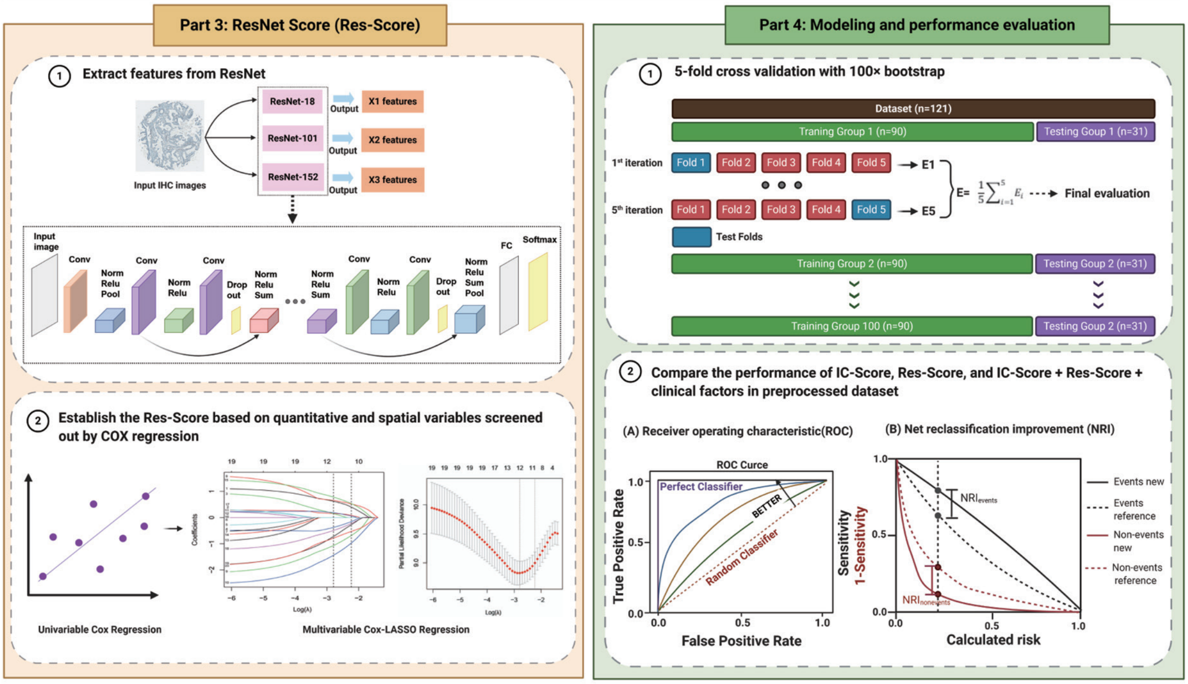


Figure 1 The research design and process of this study. We collected resected tumor tissues from 121 stage-I NSCLC patients and sliced them into paraffin sections for IHC staining of ten immune checkpoints. All sections were captured whole-slide images under microscopy (Part 1). Next, we inputted all original IHC images into the EfficientUnet-b3 model to acquire tumor cell segmentation masks. These masks were then processed with HSV thresholds, filtering, connected components extraction, and sampling to classify four types of cells in IHC images and the quantitative and spatial analysis of immune checkpoints expression on TCs and TILs. Fifty-five features were extracted upon analysis and were imputed into univariable and multivariable Cox regressions to establish the IC-Score (Part 2). Meanwhile, we also inputted all original IHC images into the Resnet to extract prognostic features and then applied univariable and multivariable Cox regressions to establish the Res-Score (Part 3). Further, to evaluate the performance of the IC-Score, Res-Score, and their combination with clinical features, we performed the AUC and NRI analysis on the dataset pretreated with 5-fold cross-validation with 100× bootstrap (Part 4). Abbreviations: NSCLC, non-small cell lung cancer; IHC, immunohistochemistry; HSV, hue, saturation, and value; TC, tumor cell; TIL, tumor-infiltrating lymphocytes; IC-Score, immune checkpoints score; Res-Score, ResNet score; AUC, area under the receiver operating characteristic curve; NRI, net reclassification index.

published (24-31) (*Figure 1*). Ten primary antibodies for KIR 2D (L1, L3, L4, S4) (BC032422/ADQ31987/NP_002246/NP036446, 1/75; Abcam, Cambridge, MA, USA), KIR 3D (L1) (AA 1-444, 1/1,500; Abcam), MHC Class II DP DQ DR (CR3/43, 1/100; Abcam), PD-1 (NAT 105, predilute; Cell Marque, Rocklin, CA, USA), PD-L1 (22C3; Dako, Carpenteria, CA, USA), GAL-9 (NBP2-45619; Novusbio, CO, USA), OX40, OX40L, LAG-3, and TIM3 (EPR4392, 1/1,000; Abcam) were applied on TMA slides from the internal cohort. The antibodies for PD-1 (ZM-0381, 1/100, Golden bridge zhongshan, Beijing, China), PD-L1 (13684S, 1/300, Cell Signaling, Beverly, MA, USA) were performed on whole IHC slides from the external cohort.

Tumor cell segmentation based on the EfficientUnet

This study performed the EfficientUnet model to segment the TCs and TILs, which was a combination of EfficientNet and UNet (32,33). UNet is a symmetric U-shaped fully convolutional neural network (CNN) developed initially for biomedical image segmentation, which processes a contraction path and an expansion path for encoder and decoder, respectively (32). EfficientNet is an adjusted CNN model which could scale the depth, width, and resolution of networks by a fixed set of scaling factors (33). Considering the better performance of low-level feature maps from encoder in the complicated spatial analysis, Baheti *et al.* originally applied EfficientNet (with intermediate low-level feature map) as the encoder of UNet (with intermediate high-level feature map) to replace the previous convolution layers (34). And the performance of EfficientUnet was much better than the other segmentation algorithms, including Dilated ResNet, ERFNet, DeepLab with ResNet18 Encoder, and the combination of UNet with ResNet or InceptionResNet (34).

As the EfficientNet has eight variants, from EfficientNet-B0 to EfficientNet-B7. According to the preliminary experiment, EfficientNet-B3 has a comparable performance and the fewest parameters compared with EfficientNet-B4 to B7. Thus, we chose the EfficientNet-B3 as the encoder, cooperated with the decoder architecture of the Unet to build a semantic segmentation network for two classes: tumor areas and non-tumor areas (*Figure 1*). The outputs of EfficientUnet were mapped to the range of 0 to 1, with a classification threshold of 0.5. There was no other preprocessing except image cropping since many diverse samples decreased the impact of color variability, with the application of data augmentation, including flip HueSaturationValue, RandomBrightness, and RandomContrast, to improve the adaptability.

We inputted all 1,859 original TMA images of 3,000×3,000 pixels (px) into the EfficientUnet-b3 model to acquire TC segmentation masks. Two pathologists labeled the tumor areas in 20 slices via the LabelMe platform (Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Boston, MA, USA; http://labelme.csail.mit.edu/Release3.0/). Further, 21,000 patches were randomly cut out of multiple sizes (100×100 px, 144×144 px, 300×300 px, or 500×500 px) as the training set and the validation set at a ratio of 8:2. Finally, the number of trainable parameters of the final EfficientUnet was approximately 13M. The technical parameters were shown as below:

Epochs: 150; batch size: 12; input size: 320×320; loss function: BCEDiceLoss; optimizer: SGD (initial learning rate: 1e−3, momentum: 0.9, weight decay: 1e−4); scheduler: cosine annealing (minimum learning rate: 1e−5, patience: 2, gamma: 2/3).

The Sørensen-Dice coefficients (dice coefficient) of the EfficientUnet model on the training and validation set were up to 0.809 and 0.829. The slider cropping was also exerted on the internal testing set to cut out multi-size patches, and the multi-size prediction results were merged to produce the results. Finally, the dice coefficient of the EfficientUnet model on the internal testing set reaches 0.793. The above experiments were conducted with Pytorch (version 1.4.0) and 3×GTX1080Ti. The representative segmentation masks obtained by the EfficientUnet model are shown in Figure S1A,B,C.

The dice coefficient has been used to quantify the similarity between the predicted segmentation mask and the ground truth, calculated as below:

$$\text{coefficient} = \frac{2|X \cap Y|}{|X| + |Y|} \qquad [1]$$

$|X|$ represented the number of pixels in the predicted segmentation mask;

$|Y|$ represented the number of pixels in the ground truth.

Classification of positive and negative cells

The segmentation task based on hue, saturation, and value (HSV) thresholding was performed on the tumor cell regions and the lymphocyte regions to distinguish the stained cells (brown) and unstained cells (blue). Here, we manually determined the threshold of positive cells on each slice by measuring the HSV of 30 cells per class with variable shades. The cutoff of each class was the range of the HSV values of the 30 cells of each class. Meanwhile, we also detected the cutoff of the impurity (such as artifacts and necroses) from 5 to 10 impure false-positive staining of each IHC image according to the same procedure as positive and negative cells. All the impure staining of each slide was excluded based on the cutoff of HSV values. Next, we repeated this procedure for each slide to define the cutoff of positive cells one by one, which could manually solve the heterogeneity of the staining intensity among different samples. Afterward, mean filtering, morphological processing, and connected components extraction and corroded into scattered points to obtain virtual sampling cells. Finally, four types of cells, including positive TCs, positive TILs, negative TCs, and negative TILs, were sampled from the extracted connected components (Figure S1A,B,C).

Calculation of the distance between cells

The density of cells (cell/mm^2) used in IHC research is an index with both spatial and quantitative information, which is calculated as the ratio of the number of positive cells and

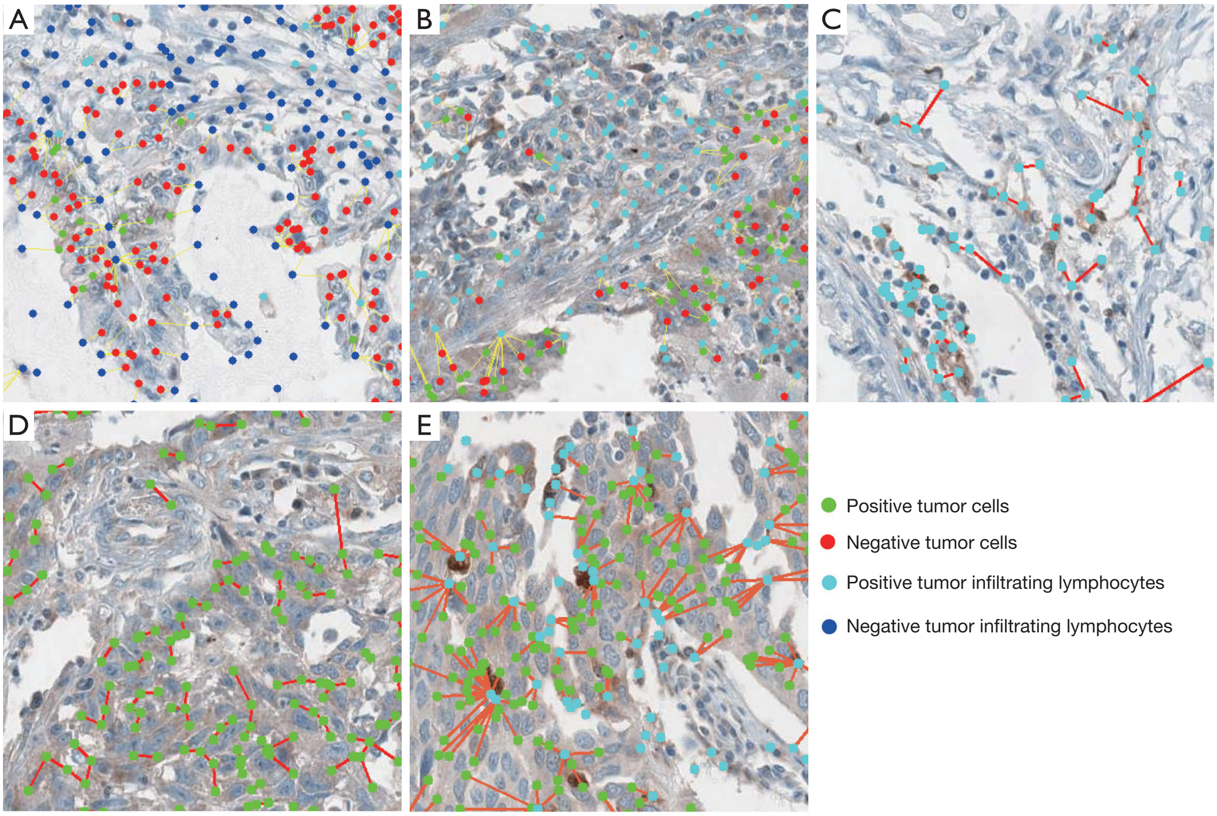

A
B
C
D
E
Positive tumor cells
Negative tumor cells
Positive tumor infiltrating lymphocytes
Negative tumor infiltrating lymphocytes

Figure 2 The representative images of segmentation and spatial analysis of the internal cohort. The local magnified images of the distance between all TCs and all TILs (A), all TCs and positive TILs (B), positive TILs and positive TILs (C), positive TCs and positive TCs (D), positive TCs and positive TILs (E). (A,B,C,D,E) were 150×150 px. Green dots represented positive TCs; red dots represented negative TCs; light blue represented positive TILs; dark blue represented negative TILs; and the red or yellow lines between cells were straight line distance between two cells. TC, tumor cell; TIL, tumor-infiltrating lymphocyte.

the size of the tissue. Since the TMAs used in the training and internal testing group were similar-sized circles, the comparative relationship between densities of different cells largely depends on the number of positive cells. To avoid the multicollinearity between the density and the number of positive cells, we detected the proximity distance between cells as spatial analysis.

The proximity distances between two types of cells in one tissue were defined as the nearest cell-to-cell distance's mean value. In this study, we detected proximity distance of TIL$_{\text{all (negative and positive)}}$ -TC$_{\text{all}}$, TC$_{\text{positive}}$-TC$_{\text{positive}}$, TIL$_{\text{positive}}$ -TC$_{\text{positive}}$, TIL$_{\text{positive}}$ -TIL$_{\text{positive}}$, and TIL$_{\text{positive}}$-TC$_{\text{all}}$ for PD-L1, galectin-9, TIM-3, OX40, OX40L, MHC-II, KID-2D, and KID-3D. Meanwhile, for PD-1 and LAG-3, only TIL$_{\text{all}}$-TC$_{\text{all}}$, TIL$_{\text{positive}}$-TIL$_{\text{positive}}$, and TIL$_{\text{positive}}$-TC$_{\text{all}}$ can be detected due to the lack of expression on tumor cells. In this study, the distances between cells were calculated in pixels (px; 200 μm =123 px). The representative calculation of the above five distances was shown in *Figure 2A,B,C,D,E*.

Construction of the immune checkpoint score (IC-Score)

According to the quantitative and spatial analysis of 1,859 TMA images, 55 parameters were extracted. Further, 55 parameters were inputted into the univariate Cox regression, and 22 significant features for overall survival (OS) and nine significant features for relapse-free survival (RFS) were screened out (*Table 1*). The least absolute shrinkage and selection operator (LASSO)-Cox regression with 10-fold cross-validation (CV) method was then performed to select significant features from the significant prognostic markers to build formulas of the

Table 1 Significant factors in univariate cox regression of OS and RFS

Factors	OS				RFS			
	No	HR	95% CI	P value	No	HR	95% CI	P value
Distance between positive TILs (one positive TIL to another TIL)								
PD-1								
Near (≤125.34 px)	27	1.000						
Far (>125.34 px)	74	1.975	1.078–3.618	0.028				
OX40L								
Near (≤48.14 px for OS, ≤61.55 px for RFS)	92	1.000			98	1.000		
Far (>48.14 px for OS, >61.55 px for RFS)	23	1.788	1.076–2.989	0.025	17	1.801	1.032–3.144	0.038
OX40								
Near (≤98.33 px)	92	1.000						
Far (>98.33 px)	26	1.752	1.078–2.846	0.024				
KIR-3D								
Near (≤13.44 px)	13	1.000						
Far (>13.44 px)	96	0.416	0.217–0.799	0.008				
Distance between positive TCs (one positive TC to another positive TC)								
TIM-3								
Near (≤546.86 px)	59	1.000						
Far (>546.86 px)	11	2.039	1.032–4.030	0.040				
galectin9								
Near (≤85.97 px for OS, ≤119.40 px for RFS)	65	1.000			75	1.000		
Far (>85.97 px for OS, >119.40 px for RFS)	35	1.820	1.105–2.999	0.019	25	1.719	1.002–2.949	0.049
KIR-2D								
Near (≤24.22 px)	68	1.000						
Far (>24.22 px)	39	1.849	1.151–2.969	0.011				
Distance between positive TCs and positive TIL (one positive TC to one positive TIL)								
OX40L								
Near (≤94.06 px)	50	1.000			50	1.000		
Far (>94.06 px)	62	0.570	0.361–0.902	0.016	62	0.576	0.366–0.906	0.017
MHC-II								
Near (≤18.68 px)	13	1.000						
Far (>18.68 px)	97	0.374	0.200–0.699	0.002				
KIR-3D								
Near (≤20.28 px)	35	1.000			35	1.000		
Far (>20.28 px)	73	2.337	1.366–3.977	0.002	73	2.444	1.379–4.330	0.002
Distance between all TCs and all TILs (one TC to one TIL)								
Near (≤44.41 px)	83	1.000						
Far (>44.41 px)	38	1.672	1.069–2.614	0.024				

Table 1 (*continued*)

Table 1 (*continued*)

Factors	OS				RFS			
	No	HR	95% CI	P value	No	HR	95% CI	P value
Distance between all TCs and positive TILs (one TC to one positive TIL)								
OX40L								
Near (≤173.56 px)	96	1.000						
Far (>173.56 px)	20	1.731	1.025–2.922	0.040				
OX40								
Near (≤445.89 px)	104	1.000						
Far (>445.89 px)	14	1.777	0.991–3.187	0.054				
MHC-II								
Near (≤44.03 px for OS, ≤135.56 px for RFS)	49	1.000			96	1.000		
Far (>44.03 px for OS, >135.56 px for RFS)	63	1.415	0.915–2.299	0.113	16	1.761	0.984–3.150	0.057
KIR-2D								
Near (≤223.71 px)	93	1.000						
Far (>223.71 px)	16	1.833	1.031–3.258	0.039				
KIR-3D								
Near (≤21.19 px)	35	1.000			35	1.000		
Far (>21.19 px)	74	2.235	1.208–3.901	0.005	74	2.017	1.181–3.443	0.010
Percentage of positive TILs								
OX40								
Low (≤44% for OS, ≤52% for RFS)	84	1.000			95	1.000		
High (>44% for OS, >52% for RFS)	24	1.739	1.087–2.781	0.026	23	2.015	1.201–3.383	0.008
galectin9								
Low (≤23% for OS, ≤20% for RFS)	22	1.000			20	1.000		
High (>23% for OS, >20% for RFS)	83	1.964	1.002–3.850	0.049	85	2.053	1.016–4.149	0.045
KIR-2D								
Low (≤37%)	46	1.000						
High (>37%)	83	0.593	0.373–0.944	0.028				
KIR-3D								
Low (≤99% for OS, ≤85% for RFS)	84	1.000			26	1.000		
High (>99% for OS, >85% for RFS)	25	0.327	0.128–0.836	0.020	83	0.327	0.128–0.833	0.019
Percentage of positive TCs								
MHC-II								
Low (≤93%)	99	1.000						
High (>93%)	13	0.425	0.185–0.979	0.044				
galectin9								
Low (≤41%)	85	1.000						
High (>41%)	20	0.450	0.223–0.909	0.026				

Table 1 (*continued*)

Table 1 (*continued*)

Factors	OS				RFS			
	No	HR	95% CI	P value	No	HR	95% CI	P value
KIR-3D								
Low (≤99%)	95	1.000						
High (>99%)	14	0.561	0.325–0.968	0.038				
Clinical factors								
Surgery type								
Wedge	2	1.000						
Segmentectomy	3	9534.267	0–3.547E+56					
Lobectomy	62	7210.961	0–2.663E+56					
Bilobectomy	7	6049.498	0–2.242E+56					
Pneumonectomy	41	14177.646	0–5.236E+56					
Sleeve lobectomy	6	10961.091	0–4.059E+56	0.020				
Pathology								
Adenocarcinoma	38	1.000						
Non-adenocarcinoma	85	1.678	1.004–2.806	0.048				
T-stage								
1	85	1.000			85	1.000		
2	36	2.011	1.260–3.210	0.003	36	2.011	1.260–3.210	0.003
N-stage								
0	63	1.000			63	1.000		
1	58	2.479	1.586–3.873	<0.0001	58	2.479	1.586–3.873	<0.0001
M-stage								
0	114	1.000			114	1.000		
1	7	4.019	1.803–8.961	0.001	7	3.855	1.734–8.574	0.001
Stage								
1A	80	1.000			80	1.000		
1B	41	5.188	3.242–8.301	<0.0001	41	4.092	2.574–6.504	<0.0001

No, number of each class; HR, hazard ratio; OS, overall survival; RFS, relapse-free survival; TC, tumor cell; TIL, tumor-infiltrating lymphocyte; KIR2D, killer cell immunoglobulin-like receptor-2D; KIR-3D, killer cell immunoglobulin-like receptor-3D; TIM-3, T cell immunoglobulin-3; LAG-3, lymphocyte activation gene-3; PD-1, programmed cell death receptor-1; PD-L1, programmed cell death ligand-1; MHC-II, major histocompatibility complex class II; OX40L, OX40-ligand. All cut-off points were determined by the X-Tile software.

IC-Score for OS and RFS (*Figure 1*).

Construction of the ResNet score (Res-Score)

The ResNet models we used in this study were ResNet 18, ResNet 101, and ResNet 152 in Pytorch 1.4.0, which were pre-trained on ImageNet (Table S1). All IHC images were inputted into the ResNet (35) to extract prognostic features. We calculated the mean of raw features obtained from all IHC images for the same patient. Further, we applied the univariable Cox regression and multivariable LASSO-Cox regression on these raw features to establish the Res-Score

(*Figure 1*).

Construction of the integrated score

To optimize the IC-Score and Res-Score performance, we inputted all significant clinical and IHC analysis features (listed in *Table 1*) with the significant features extracted via the ResNet into the LASSO-Cox regression mentioned above to build formula of the integrated score (*Figure 1*).

Validation from the external cohort and public databases

To validate the accuracy of quantitative and spatial analysis of four types of cells and the performance of Res-Score, we recruited an external cohort of 30 NSCLC patients with 214 whole-slide images at Shanghai Pulmonary Hospital. Considering the large size of raw whole-slide images (>50,000×50,000 px), five represented regions of interest (ROIs) of each slide were selected by two pathologists, with a size of 1,238×849 px. All these ROIs were inputted into the EfficientUnet trained by the internal training groups, and the same pre-trained ResNet was used in the internal cohort.

The external validation of the prognostic role and correlation of immune checkpoints proteins were performed via the survival plot and correlation analysis section from Gene Expression Profiling Interactive Analysis (GEPIA) database within lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) samples (http://gepia.cancer-pku.cn/). The survival analysis between different mRNA levels of PD-1/PD-L1 was from the pathology section of the Human Protein Atlas (HPA) database within LUAD and LUSC samples (https://www.proteinatlas.org/). The interaction analysis of correlated proteins in the internal cohort was validated via the multiple proteins section from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING; https://string-db.org/). Moreover, to enhance the prognostic signature of PD-1/PD-L1, we introduced a widely reported systemic inflammatory index, the neutrophil-to-lymphocyte ratio (NLR), in the external cohort. NLR was calculated as the ratio of neutrophils and lymphocytes count in the pre-operate blood routine results.

Statistical analysis

The multiple imputation (36,37) was performed to fill the missing values. The clinical outcomes in this study were OS and RFS. The OS time was determined from the surgery to the death induced by any cause, while the RFS time was determined from the surgery to the disease recurrence.

Bivariable association between predictive variables and OS or RFS was evaluated by the Cox proportional hazards model and the log-rank test. The multivariable LASSO-Cox regressions were used to establish integrated predictive models for OS and RFS. Further, the performance of risk-stratification was assessed by the net reclassification index (NRI) and time-dependent receiver operating characteristic curve (ROC). The modeling process and predictive accuracy evaluation were exerted via 5-fold CV with 100× bootstrap resampling.

The optimal cutoff values of each feature were determined by the X-Tile software (38), which is based on the minimum P value or maximum Chi-square value. The one-way analysis of variance (ANOVA) was performed for statistical significance. All data analysis in this study was accomplished by R software (version 3.5.0, R Core Team), Python (version 3.7, Python Software Foundation), and GraphPad Prism (version 8.0, GraphPad Software). Statistical tests were two-sided, and P<0.05 was considered statistically significant.

Results

Characteristics of the patient cohort

The internal cohort included 1,859 TMA images of 121 NSCLC patients. In this cohort, 96 (79.3%) patients were males, and 117 (96.7%) patients were smokers. Most patients (n=92, 76.0%) were under 70 years. Eighty (66.1%) patients had stage-IA NSCLC, while the rest had stage-IB NSCLC. Thirty-six (29.8%) patients had LUAD, and the rest had other NSCLC (Table S2).

The external set contained 214 IHC images (with a choice of 5 representative ROIs) on PD-1 and PD-L1 from 30 resected NSCLC patients. In this cohort, 73.3% were under 70 years (n=22), 73.3% were male (n=22), 30.0% were smokers (n=9), and 66.7% were diagnosed as adenocarcinoma (n=20). Moreover, 43.3% of the external cohort were at stage-I disease (n=13), while 33.3% (n=10) and 23.3% (n=7) were diagnosed as stage-II and stage-III disease. In short, the external cohort had a similar characteristic of age and sex with the internal cohort, except the surgery procedures, TNM-stage, smoking status, and histology (Table S1).

Quantitative and spatial analysis of 10 immune checkpoints

As mentioned in the Methods, we performed the EfficientUnet to segment TCs and TILs (Figure S1A,B,C and *Figure 2A,B,C,D,E*). Among all immune checkpoints, TIM-3 and PD-L1 presented a relatively farther distance between TC$_{\text{positive}}$-TC$_{\text{positive}}$ (TIM-3: 306.0±39.70 px; PD-L1: 240.0±46.29 px), TIL$_{\text{positive}}$-TIL$_{\text{positive}}$ (TIM-3: 131.9±16.97 px; PD-L1: 108.0±18.05 px), TC$_{\text{all}}$-TIL$_{\text{positive}}$ (TIM-3: 283.1±25.03 px; PD-L1: 386.9±41.14 px), and TC$_{\text{positive}}$-TIL$_{\text{positive}}$ (TIM-3: 152.4±19.51 px; PD-L1: 152.7±32.60 px). Moreover, KIR-3D appeared as the most densely distributed marker in the TME, with the nearest distance between TC$_{\text{positive}}$-TC$_{\text{positive}}$ (50.60±9.862 px), TIL$_{\text{positive}}$-TIL$_{\text{positive}}$ (24.33±3.318 px), TC$_{\text{all}}$-TIL$_{\text{positive}}$ (44.98±6.256 px), TC$_{\text{positive}}$-TIL$_{\text{positive}}$ (37.08±4.834 px). The other spatial features were presented in Figures S2-S4, while the original distribution of the distance between TIL$_{\text{positive}}$-TC$_{\text{all}}$ was shown in Figure S5A,B,C,D,E,F,G,H,I,J,K.

The percentage of positive TCs or TILs was calculated by the ratio of the number of positive TCs (or TILs) and all TCs (or TILs) on whole sections. KIR-3D presented the largest proportion of positive cells in both TCs (85.77%±2.316%) and TILs (75.50%±2.352%), while TIM-3 and PD-L1 had the lowest positive percentage of TCs (TIM-3: 4.712%±0.8868%; PD-L1: 11.83%±2.129%) and TILs (TIM-3: 12.28%±1.806%; PD-L1: 19.38%±2.721%). The original distribution and mean values of other quantitative features were presented in Figure S5L,M. Moreover, we compared the AI-based quantitative results with manual evaluation from two pathologists and obtained a strong correlation (all R: 0.8476–0.9335; *Figure 3A*). Thus, this model presented a labor-saving way to automatically identify four types of cells with a comparable accuracy with manual recognition, which may promote the clinical routine test of multiple immune checkpoints.

Pearson correlation analysis identified a moderate correlation among LAG-3, OX40, OX40L, and KIR2D (R>0.5). As expected, the percentage of TIL$_{\text{OX40+}}$ and TC$_{\text{OX40L+}}$ (R=0.7236; Figure S5N), and the percentage of TIL$_{\text{OX40L+}}$ and TC$_{\text{OX40+}}$ (R=0.7294; Figure S5O) presented a relatively higher correlation. Surprisingly, the percentage of TIL$_{\text{KIR2D+}}$ (R=0.5737; Figure S5P) and TIL$_{\text{OX40+}}$ (R=0.5720; Figure S5Q) were significantly correlated with that of TIL$_{\text{LAG-3+}}$. The percentage of TC$_{\text{OX40+}}$ (r=0.5654; Figure S5R) and TC$_{\text{OX40L+}}$ (R=0.5564; Figure S5S) both had a similar correlation with TC$_{\text{LAG3+}}$.The distance of TC$_{\text{OX40L+}}$-TC$_{\text{OX40L+}}$ and the distance of TIL$_{\text{2D+}}$-TIL$_{\text{2D+}}$ were the only pair of spatial variables with significant correlation (R=0.628; Figure S5T). In conclusion, the interaction among the above three pathways (MHC-II/LAG-3, OX40/OX40L, and KIR2D) revealed the great potential of combining immune checkpoint inhibitors, which could provide new ideas for clinical combinational immunotherapy.

The impact of quantitative and spatial analysis of immune checkpoints on prognosis

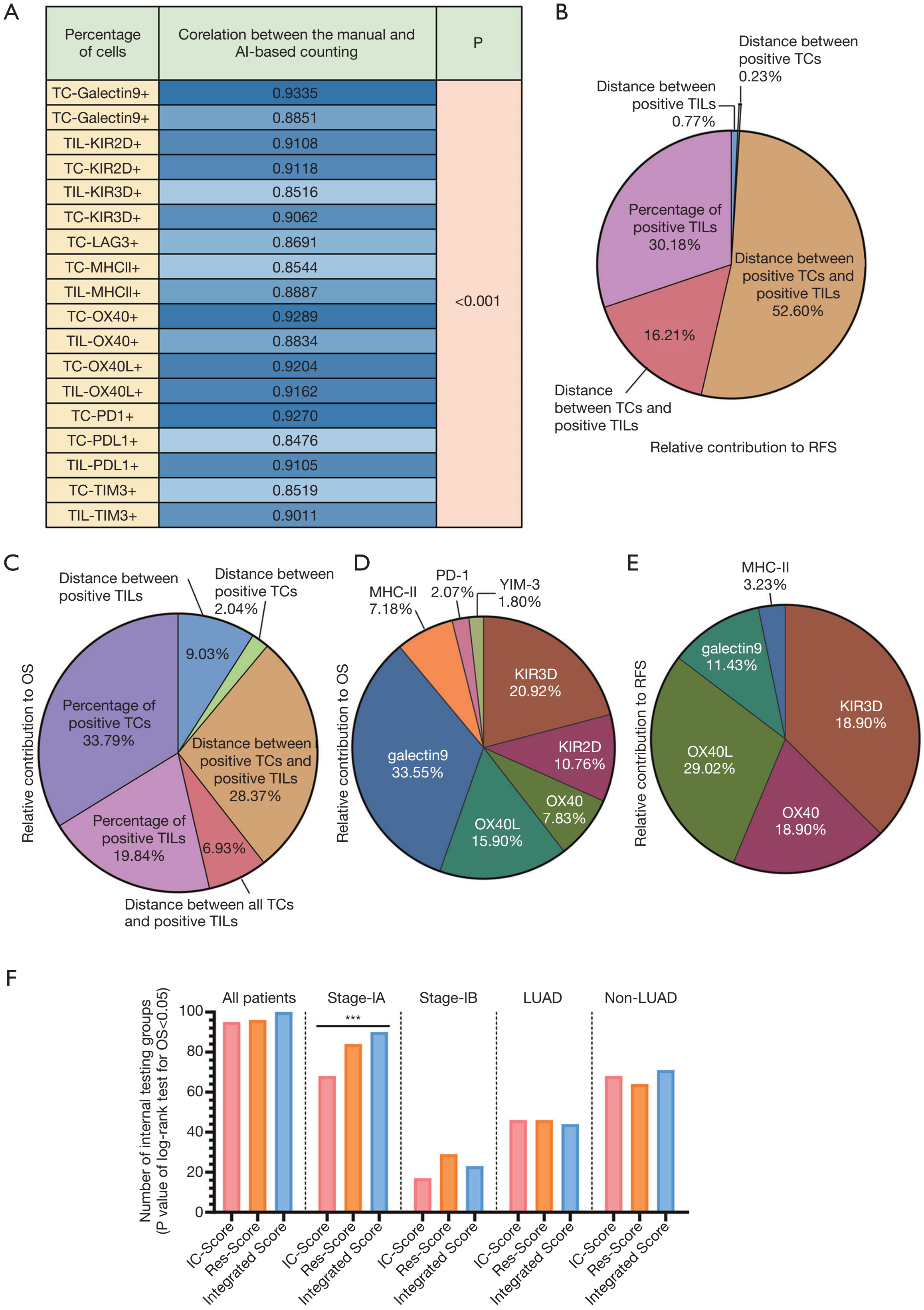
When performing the univariate Cox regression analysis on the OS based on the raw data, 20 IHC-related variables remained significant risk factors (P<0.05). Moreover, we assumed the distance between TC$_{\text{all}}$-TIL$_{\text{OX40+}}$ (P=0.054) as a marginally significant factor for OS. Among these 21 variables, the distance between TC$_{\text{all}}$-TIL $_{\text{positive}}$ (n=4), the distance between TIL$_{\text{positive}}$-TIL$_{\text{positive}}$ (n=4), and the percentage of positive TILs (n=4) led in quantity. For ten markers included in this study, KIR-3D-related variables presented an enormous amount (n=5), while PD-L1 or LAG-3 related variables did not prove independent prognostic factors (*Table 1* and Table S2).

In terms of RFS, eight variables were independently significant (P<0.05), and the distance of TC$_{\text{all}}$-TIL$_{\text{MHCII+}}$ (P=0.057) was the only marginal risk factor (*Table 1*; Table S2). Notably, all independent prognostic factors for RFS (P<0.05) were a subset from those for OS.

Construction of the IC-Score

As mentioned in Methods, we performed data preprocessing via the multiple imputations to complete all 121 patients' variables and then randomly split all patients into 100 training groups (n=90) and 100 internal testing groups (n=31).

To establish an IC-Score, we inputted all significant prognostic factors for OS and RFS (listed in *Table 1*) into a multivariable LASSO-Cox regression based on the minimal lambda (39), including two marginally significant variables. When analyzing the parameter's contribution in the IC-Score for predicting OS, the percentage of positive TCs, the distance between positive TCs and TILs, and the percentage of positive TILs were the most significant classification (*Figure 3B*). However, the percentage of TCs were not substantial for RFS (*Figure 3C*). Moreover, the



Percentage of cells	Corelation between the manual and AI-based counting	P
TC-Galectin9+	0.9335	<0.001
TC-Galectin9+	0.8851	
TIL-KIR2D+	0.9108	
TC-KIR2D+	0.9118	
TIL-KIR3D+	0.8516	
TC-KIR3D+	0.9062	
TC-LAG3+	0.8691	
TC-MHCII+	0.8544	
TIL-MHCII+	0.8887	
TC-OX40+	0.9289	
TIL-OX40+	0.8834	
TC-OX40L+	0.9204	
TIL-OX40L+	0.9162	
TC-PD1+	0.9270	
TC-PDL1+	0.8476	
TIL-PDL1+	0.9105	
TC-TIM3+	0.8519	
TIL-TIM3+	0.9011	

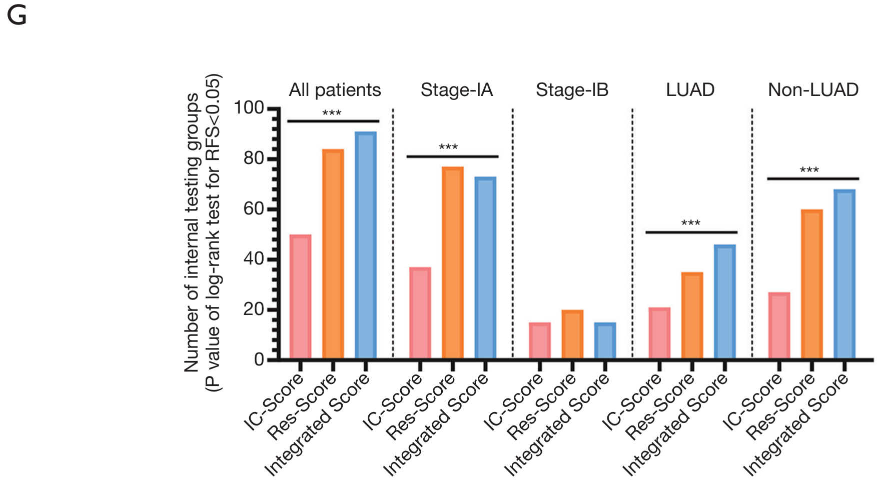


Figure 3 The composition and function analysis of IC-Score, Res-Score, and integrated score. (A) The correlation coefficient (R value) between the manual and AI-based counting via the Spearman correlation test. The relative contribution of each quantitative and spatial classification in the IC-Score for OS (B) and RFS (C) is based on all internal testing groups'' mean coefficient. The relative contribution of each immune checkpoint in the IC-Score for OS (D) and RFS (E) is based on all internal testing groups'' mean coefficient. The number of internal testing groups with a P- value <0.05 in the log-rank test for OS (F) and RFS (G). ***P<0.001. From left to right are sub-group analysis in all patients, stage-IA disease, stage-IB disease, LUAD, and non-LUAD. TC, tumor cell; TIL, tumor-infiltrating lymphocyte; KIR2D, killer cell immunoglobulin-like receptor-2D; KIR-3D, killer cell immunoglobulin-like receptor-3D; TIM-3, T cell immunoglobulin-3; LAG-3, lymphocyte activation gene-3; PD-1, programmed cell death receptor-1; PD-L1, programmed cell death ligand-1; MHC-II, major histocompatibility complex class II; OX40L, OX40-ligand; OS, overall survival; RFS, relapse-free survival.

spatial parameters made an outstanding contribution, with a proportion of 46.36% and 69.82% for OS and RFS, respectively. Further, the spatial and quantitative parameters of KIR-3D were a notable predictor, which occupied a second-highest proportion followed by galectin-9 in OS and the highest proportion in RFS (*Figure 3D*,*E*).

When exerting Kapan-Meier survival curves with the log-rank test on 100 internal testing groups, patients with a low IC-Score from 95 groups had a significantly longer OS time than the ones with a high IC-Score. In subgroup analysis, the OS time of stage-IA and stage-IB could be significantly stratified by the IC-Score in 68 internal testing groups and 17 internal testing groups, respectively (stage IA *vs.* stage IB: P<0.001). As for the pathological subtype, the OS time of LUAD patients and non-LUAD patients in 46 and 68 internal groups could be stratified with the IC-Score (LUAD *vs.* non-LUAD: P=0.0017; *Figure 3F*).

In 50% (50/100) of internal testing groups, patients with a higher IC-Score were associated with a higher risk of recurrence. In stage-IA and stage-IB, patients from 37 groups and 15 groups were fit with the association of Res-Score and RFS (stage IA *vs.* stage IB: P<0.001). Patients with a LUAD from 21 groups and those with a non-LUAD from 27 groups had a different RFS time with a different IC-Score (LUAD *vs.* non-LUAD: P=0.3205; *Figure 3G*). The survival analysis of OS and RFS in three training groups and internal testing groups were shown in *Figure 4* and Figure S6, respectively.

Construction of the Res-Score

As for the construction of Res-Score, we imputed all significant features extracted by the ResNet (Table S3) into a multivariable LASSO-Cox regression. In 96 internal testing groups, patients with a low Res-Score presented with a dramatically longer OS time than those with a high Res-Score. Among patients with stage-IA of 84 internal testing groups, the Res-Score was still a predominant marker to distinguish OS time. A similar conclusion can be drawn for patients in stage-IB from 29 internal testing groups (stage-IA *vs.* stage-IB: P<0.001). The risk-stratification function of the Res-Score was also demonstrated in patients with a LUAD of 46 internal testing groups and those with a non-

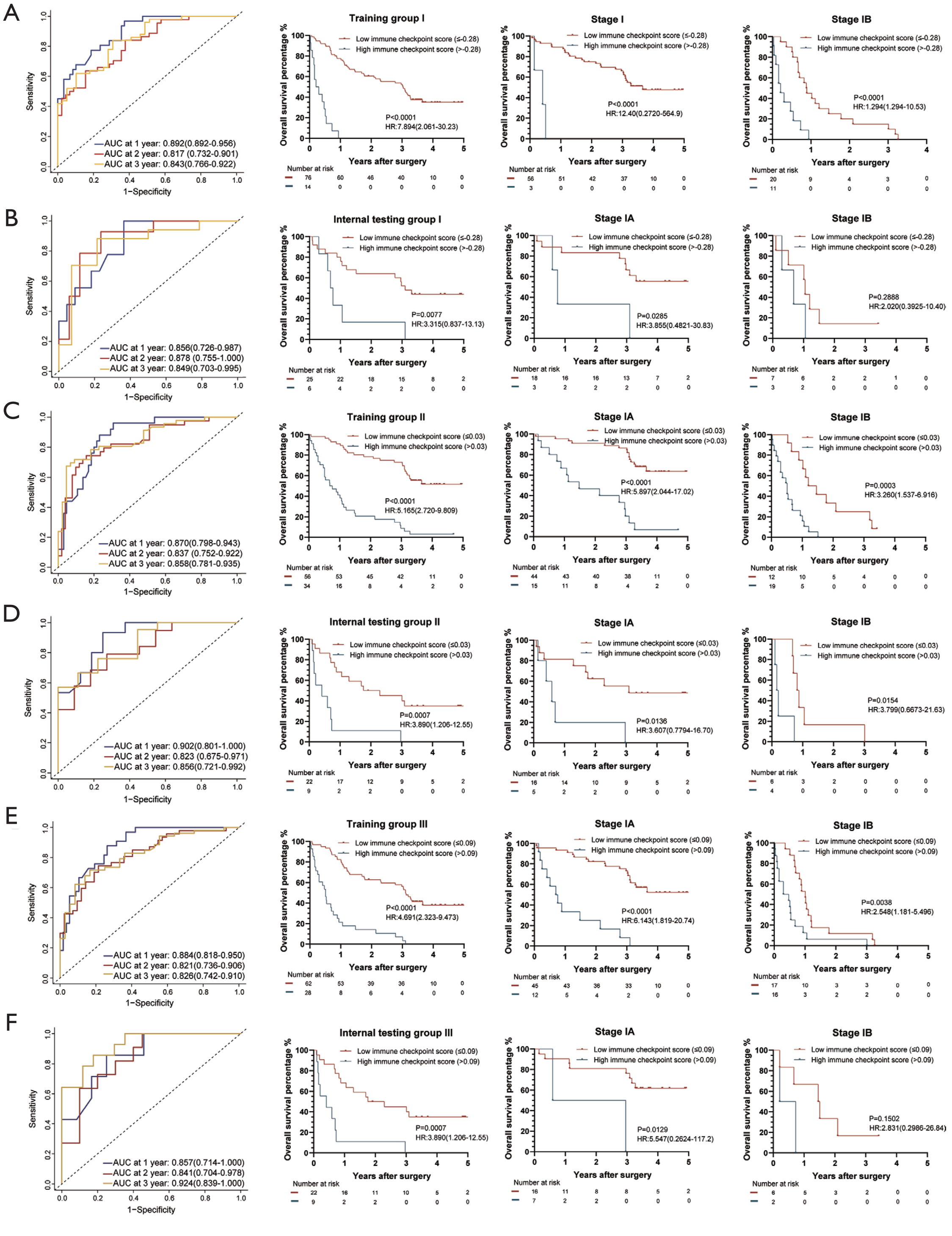


Figure 4 The IC-Score for OS measured by time-dependent ROC curves and Kaplan-Meier survival in the representative 3 training and internal testing groups. (A) Training group I; (B) internal testing group I; (C) training group II; (D) internal testing group II; (E) training group III; (F) internal testing group III. We used AUCs at 1, 2, and 3 years to assess prognostic accuracy of OS, and calculated P values using the log-rank test. Data represent AUC or P value. HR, hazard ratio; AUC, area under ROC; ROC, receiver operator characteristic. The cut-off point was determined by the X-Tile software, and the time-dependent AUC with 95% CI was calculated by the "timeROC" package of R software.

LUAD from 64 internal testing groups (LUAD *vs.* non-LUAD: P=0.0105; *Figure 3F*). The survival analysis of OS in three training groups and internal testing groups were shown in *Figure 5*.

A high Res-Score is correlated to a growing risk of recurrence in 84 internal testing groups for all patients. In patients with stage-IA and stage-IB, 77 groups and 20 groups met the above rule (stage-IA *vs.* stage-IB: P<0.001). In patients with a LUAD and a non-LUAD, 35 groups and 60 groups met the above rule (LUAD *vs.* non-LUAD: P<0.001; *Figure 3G*). The survival analysis of RFS in three training groups and internal testing groups were shown in Figure S7. In summary, the IC-Score distinguished OS better than RFS, while the Res-Score had a more stable prognostic capability between OS and RFS. Moreover, both the IC-Score and the Res-Score were less significant in stage-IB and non-LUAD, which might attribute to the small sample of stage-IB or non-LUAD.

Construction of the integrated score based on the IC-Score and the Res-Score

We further combined IC-Score, Res-Score, and clinical factors to establish an integrated score. Here, we input all significant features from the univariable analysis into a multivariable LASSO-Cox regression, including clinical characteristics, spatial or quantitative features of immune checkpoints (listed in *Table 1*), and the prognostic features extracted from ResNet models.

Surprisingly, the integrated score was a robust predictor of OS in all internal testing groups, where patients with a high integrated score presented with a higher probability of death. According to the subgroup analysis, 90% of internal testing groups in stage-IA and 23% of stage-IB demonstrated a similar conclusion (stage-IA *vs.* stage-IB: P<0.001). Forty-four groups of patients with a LUAD and 71 groups of those with a non-LUAD presented a significantly lower OS rate when classified as a high integrated score (LUAD *vs.* non-LUAD: P<0.001; *Figure 3F*). The survival analysis of OS in three training groups and internal testing groups were shown in *Figure 6*.

Patients with a high integrated score were more likely to relapse in 91% of internal testing groups for all stages. In stage-IA, patients' RFS time from 73% of internal testing groups could be dramatically divided into two classifications. In stage-IB, 15% of internal testing groups demonstrated a similar situation (stage-IA *vs.* stage-IB: P<0.001). The integrated score can risk-stratify patients' RFS time with a LUAD and non-LUAD in 46 and 68 internal testing groups, respectively (LUAD *vs.* non-LUAD: P=0.0017; *Figure 3G*). The survival analysis of RFS in three training groups and internal testing groups were shown in Figure S8. In summary, the predictive efficiency of the integrated score for OS was also more likely to be significant than those for RFS. Meanwhile, the performance of the integrated score also correlated with TNM-stage and pathological subtypes. Moreover, the prognostic significance of the integrated score for RFS, especially in pathological subgroups, was superior to the IC-Score and Res-Score (*Figure 3G*).

Performance of IC-Score, Res-Score, and integrated score

The ability of the integrated score to predict 1-year [mean area under the receiver operating characteristic curve (AUC): 0.907], 2-year (mean AUC: 0.913), and 3-year OS (mean AUC: 0.892) was superior to that of IC-Score, Res-Score, clinical factors, and their combinations (*Figure 7A,B,C*; all P<0.001). Moreover, the integrated score was proved as the most potent predictor for 1-year (mean AUC: 0.854), 2-year (mean AUC: 0.864), and 3-year RFS (mean AUC: 0.843; *Figure 7D,E,F*; all P<0.001). However, the predictive performance of IC-Score (1-year mean AUC: 0.713; 2-year mean AUC: 0.697; 3-year mean AUC: 0.689) for relapse was poorer than that of the Res-Score (1-year mean AUC: 0.781; 2-year mean AUC: 0.816; 3-year mean AUC: 0.804) and clinical features (1-year mean AUC: 0.774; 2-year mean AUC: 0.735; 3-year mean AUC: 0.691). Thus, the performance of the combination of the Res-Score with clinical features (1-year mean AUC: 0.843; 2-year mean AUC: 0.863; 3-year mean AUC: 0.843) was extremely close to that of the integrated score.

Due to the evenly matched performance of the integrated score and the Res-Score combination with clinical features, we further performed the NRI analyses (*Figure 7G*). Most of the NRI analysis results were in accordance with that of ROC analysis. Compared with other models, the integrated score presented a notable improvement of predictive accuracy, except the Res-Score combined with clinical variables for RFS. Although the integrated score demonstrated a weak advantage on predicting 1-year (mean NRI: 0.028) and 2-year RFS (mean NRI: 0.010), the predictive accuracy of the integrated score for 3-year RFS was marginally more insufficient than that of the combination of the Res-Score with clinical variables (mean NRI: –0.005). In summary, the integrated score, based on quantitative and spatial analysis of immune

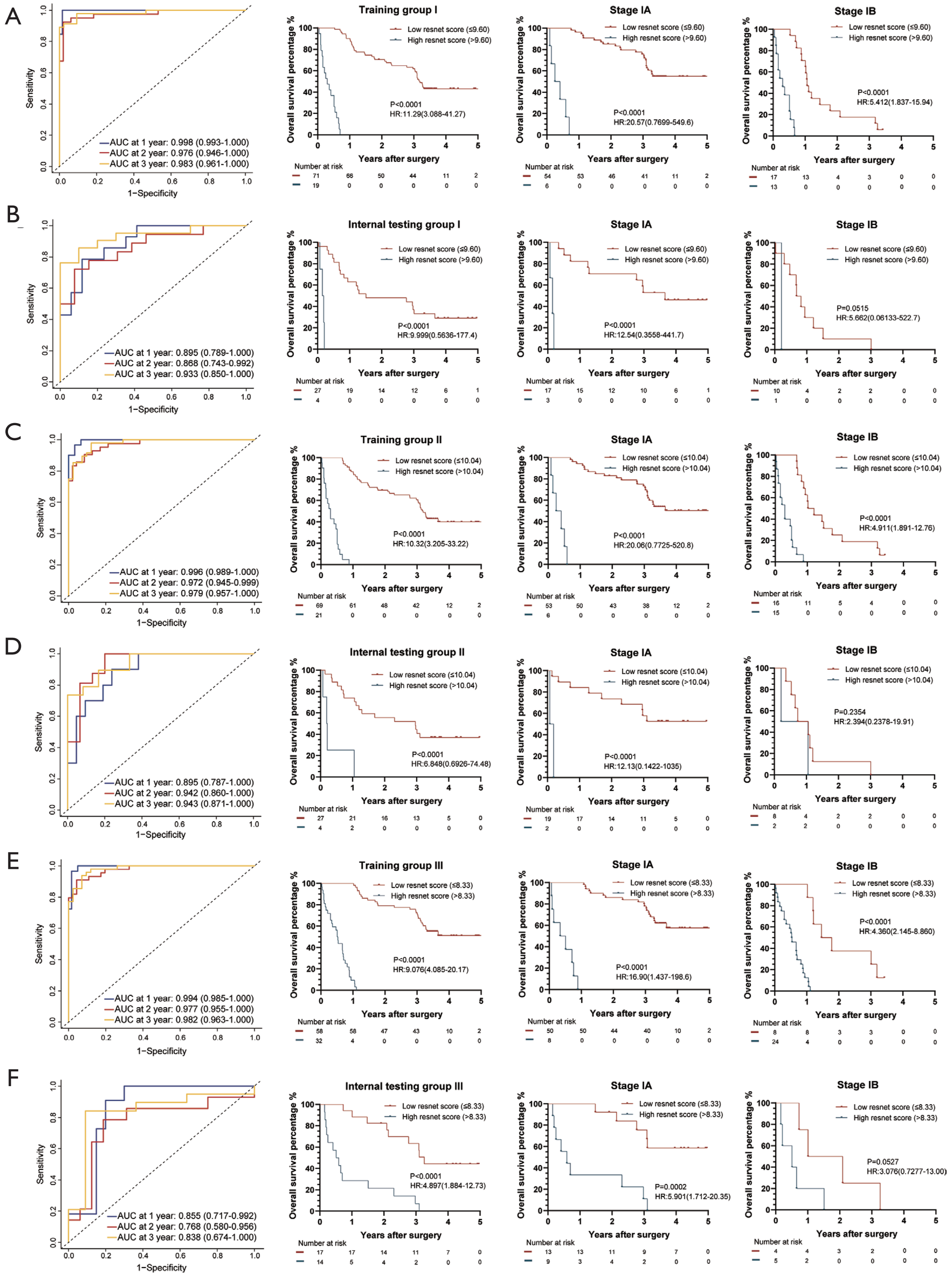


Figure 5 The Res-Score for OS measured by time-dependent ROC curves and Kaplan-Meier survival in the representative 3 training and internal testing groups. (A) Training group I; (B) internal testing group I; (C) training group II; (D) internal testing group II; (E) training group III; (F) internal testing group III. We used AUCs at 1, 2, and 3 years to assess prognostic accuracy of OS, and calculated P values using the log-rank test. Data represent AUC or P value. HR, hazard ratio; AUC, area under ROC; ROC, receiver operator characteristic. The cut-off point was determined by the X-Tile software, and the time-dependent AUC with 95% CI was calculated by the "timeROC" package of R software.

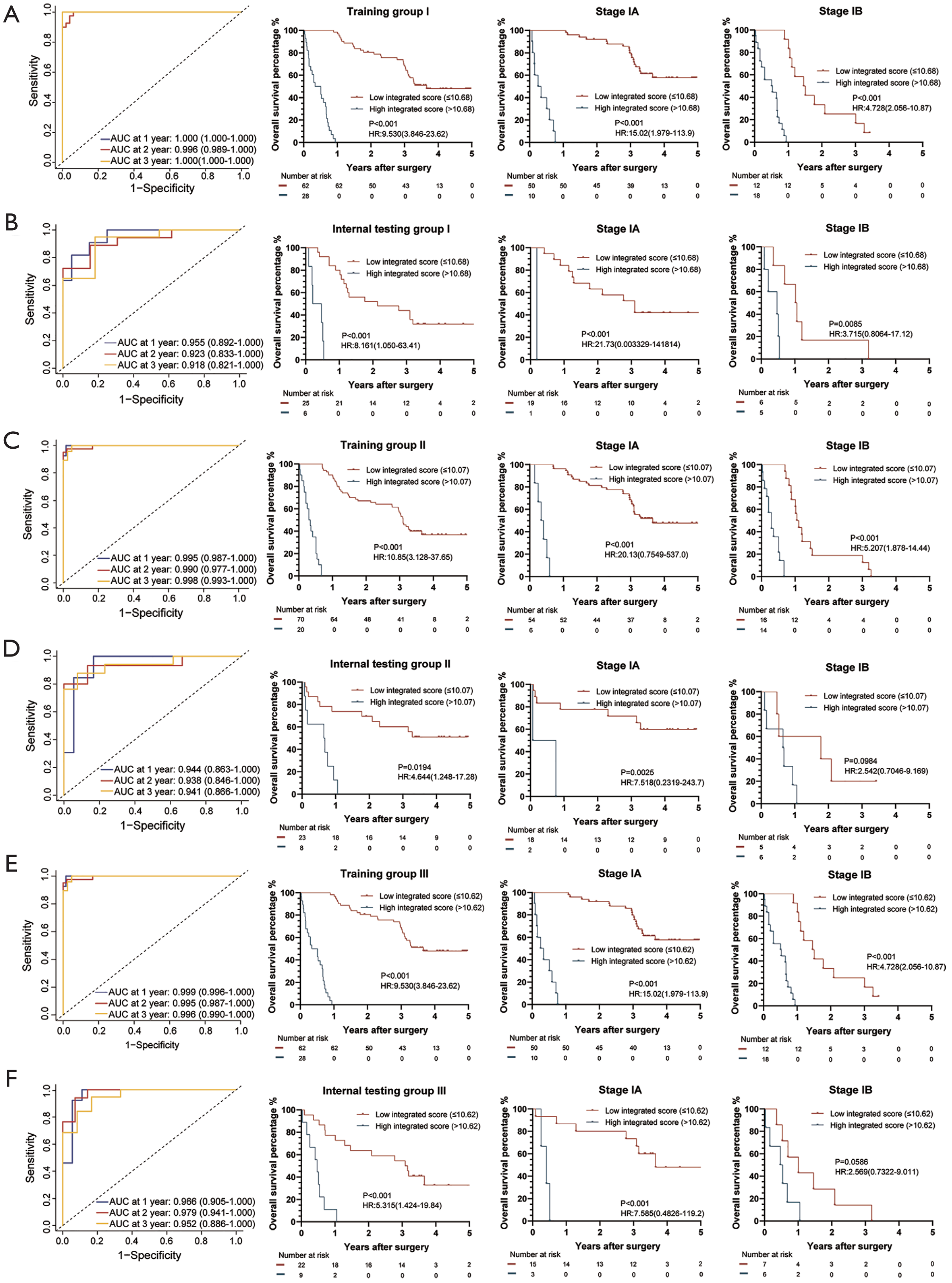
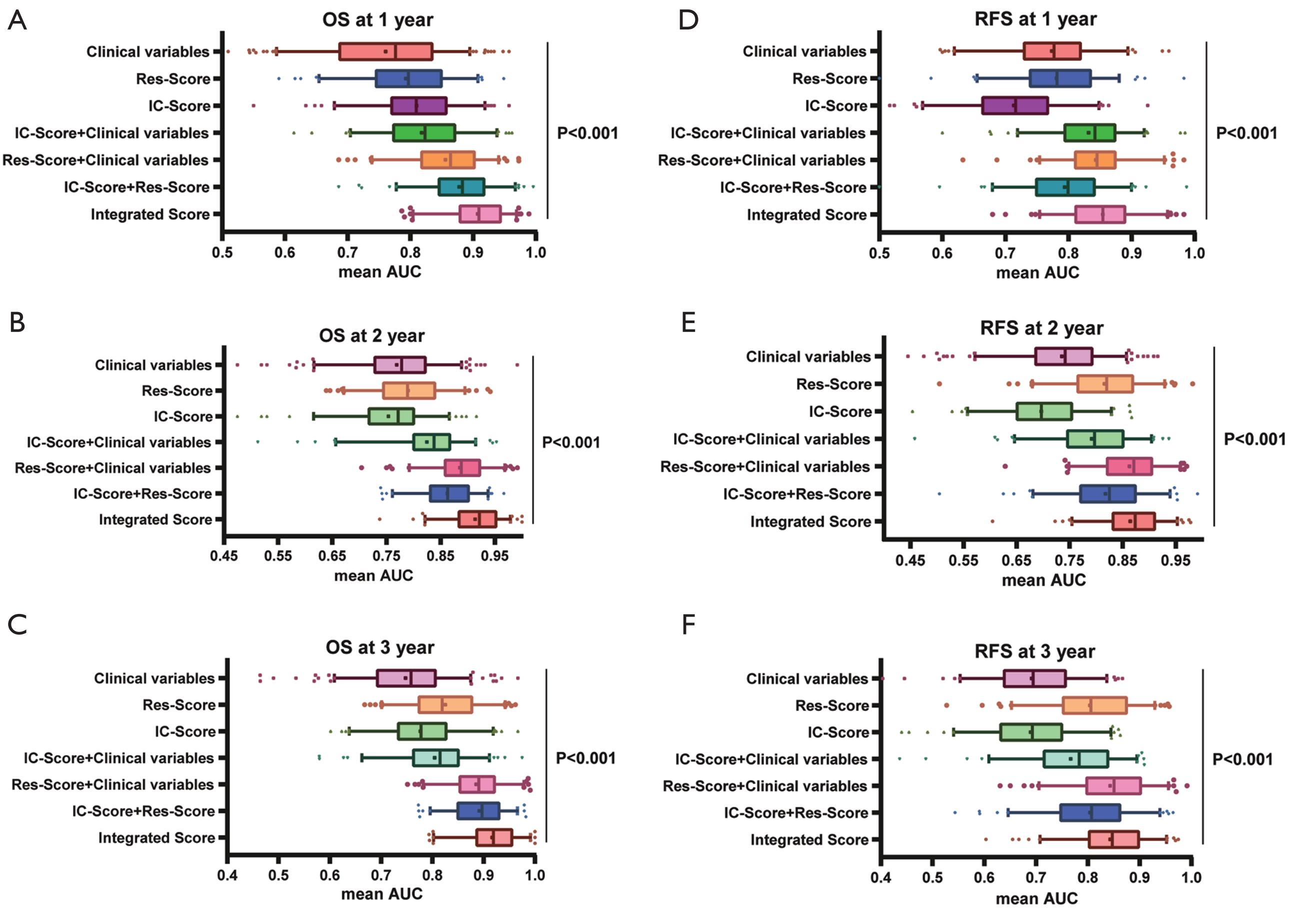


Figure 6 The integrated score for OS measured by time-dependent ROC curves and Kaplan-Meier survival in the representative 3 training and internal testing groups. (A) Training group I; (B) internal testing group I; (C) training group II; (D) internal testing group II; (E) training group III; (F) internal testing group III. We used AUCs at 1, 2, and 3 years to assess prognostic accuracy of OS, and calculated P values using the log-rank test. Data represent AUC or P value. HR, hazard ratio; AUC, area under ROC; ROC, receiver operator characteristic. The cut-off point was determined by the X-Tile software, and the time-dependent AUC with 95% CI was calculated by the "timeROC" package of R software.



G

New model vs. Old models		mean NRI of 100 testing groups					
New model	Old models	OS at 1 year	OS at 2 year	OS at 3 year	RFS at 1 year	RFS at 2 year	RFS at 3 year
Integrated Score	Res-Score	0.207	0.212	0.157	0.121	0.071	0.046
	IC-Score	0.185	0.271	0.234	0.226	0.253	0.258
	Clinical variables	0.281	0.185	0.272	0.088	0.168	0.162
	IC-Score+Res-Score	0.071	0.102	0.055	0.095	0.067	0.048
	IC-Score+Clincal variables	0.175	0.165	0.192	0.033	0.093	0.107
	Res-Score+Clincal variables	0.104	0.063	0.063	0.028	0.010	-0.005

Figure 7 The performance evaluation of all prognostic models. The predictive accuracy for 1-year OS (A), 2-year OS (B), 3-year OS (C), 1-year RFS (D), 2-year RFS (E), and 3-year RFS (F) based on the AUC with 100× bootstrap resampling for each model is shown in a box plot. Median values of 100× bootstrap resampling are shown with thick lines, and the mean values are shown by dots in the boxes. The mean NRIs of the comparison between the integrated-score and the rest models based on 100× bootstrap resampling is presented in table (G). OS, overall survival; RFS, relapse-free survival; IC-Score, immune checkpoint score; Res-Score, ResNet score; AUC, AUC, area under the receiver operating characteristic curve.

checkpoints, and prognostic features extracted by ResNet models, demonstrated a stable potential to predict OS and RFS in resected NSCLC. The solid prognostic value of the integrated score provided an approach to a convenient risk-stratification of the patient by inputting all patients' relevant immune-checkpoint-staining IHC images into the model.

External validation of the EfficientUnet model and the Res-Score

In the internal cohort, the percentage of the cells expressed galectin9, OX40, OX40L, KIR-2D, and KIR-3D played an essential function in OS or RFS. According to the GEPIA dataset, the gene expression of KIR2DL1 (P=0.029 for

OS), KIR2DL3 (P=0.017 for OS; P=0.038 for RFS), and KIR2DL4 (P=0.014 for OS) were significant for survival in LUAD and LUSC patients, with a similar hazard ratio (HR) of the internal cohort (Figure S9A,B,C). Although the survival time with different expression levels of the rest molecules did not reach a statistical difference, their HR trends were also consistent with internal groups (Figure S9D,E,F,G,H). Moreover, the correlation of LAG3-OX40 (P<0.0001, R=0.39), LAG3-OX40L (P<0.0001, R=0.44), LAG3-KIR2D (KIR2D-L1: R=0.37; KIR2D-L3: R=0.48; KIR2D-L4: R=0.59; KIR2D-S4: R=0.34; all P<0.0001), except the correlation of KIR2D-OX40 (all R≤0.25), were also validated on gene-level from the GEPIA dataset (Figure S10). The interaction of KIR2D and OX40 proteins was relatively indirect and mediated by LAG3 from the STRING database, which might be postulated to explain the lower correlation of KIR2D and OX40 (Figure S11A). In short, these similar trends between the public data and the internal cohort validated the reliability of the conclusion, which were drawn based on the segmentation of TCs and TILs and further classifications of positive and negative cells.

According to the HPA dataset, the mRNA levels of PD-L1 and PD-1 were not significant for OS (Figure S11B,C), which is the same as the conclusion from the internal cohort. With the quantitative and spatial analysis of PD-1/PD-L1 expressions through the EfficientUnet in the external cohort (*Figure 8A,B*), the principal component analysis (PCA) was further exerted to obtain a 5-dimension PD-1/PD-L1 signature whose cumulative proportion of explained variance approached 76.2% (Figure S11D). The distance of TIL_{PD1+}-TIL_{PD1+}, TIL_{PDL1+}-TIL_{PDL1+}, TC_{PDL1+}-TIL_{PDL1+}, and TCs-TILs were the features with the most representation (cos2; *Figure 8C*) and contribution (contrib; Figure S11E). The image processing of IHC images on PD-1/PD-L1 from the external testing cohort was presented in the Figure S12. Further, 30 patients were clustered into two groups through the k-mean clustering, while the two clusters did not present any significant difference in OS and RFS (*Figure 8D,E*). Surprisingly, the combination of preoperative NLR from the blood routine and the PD-1/PD-L1 signature was not a robust prognostic index for OS (*Figure 8F*), but was vital for RFS (P<0.0001 for RFS; *Figure 8G*), although NLR was not a significant feature for OS and RFS in this population (*Figure 8H,I*). Thus, the combination of PD-1/PD-L1 signature with NLR might be a potential prognostic biomarker of clinical immunotherapy, compared with the PD-1/PD-L1 signature alone.

Moreover, we input all ROIs of the external cohort into the ResNet trained by the internal cohort to obtain raw features. Further, we took the corresponding features into the internal cohort formula to calculate the Res-Score for the external cohort. The patients with a higher Res-Score presented a significantly lower OS time and RFS time than those with a lower Res-Score (P<0.001 for OS; P=0.0097 for RFS; *Figure 8J,K*). The Res-Score also presented a stable predictive ability for OS [1.5-year AUC: 0.800 (0.622–0.978); 1.75-year AUC: 0.868 (0.718–1.019); 2-year AUC: 0.861 (0.703–1.019)] and RFS [1.5-year AUC: 0.875 (0.667–1.083); 1.75-year AUC: 0.941 (0.826–1.057); 2-year AUC: 0.826 (0.627–1.026)] (*Figure 8L*). In short, the high performance of the Res-Score in the external cohort validated its generalization ability in various populations, which provided the great potential to assist clinical decisions in various institutes.

Discussion

This study implemented quantitative and spatial analysis of ten immune checkpoints on TCs and TILs based on the EfficientUnet and ResNet to establish predictors for OS and RFS. Further, the IC-Score and Res-Score constructed by LASSO-Cox regressions were significant prognostic biomarkers for OS and RFS (all mean AUC >0.75). The integrated score was a combination of the IC-Score, Res-Score, and clinical variables, demonstrating a notable improvement in prognostic ability. Moreover, the prognostic role and correlation of significant immune checkpoint proteins were validated from public datasets. Further, the Res-Score demonstrated a stable performance in the external cohort, presenting a generalization ability among different populations. In conclusion, we revealed that the spatial analysis and deep learning of single-plex chromogenic IHC held an excellent value in risk-stratify relapse and death in resected NSCLC.

Current studies have confirmed that the PD-1/PD-L1 axis was prevalent in resected NSCLC (stage I-III) by identifying a tangible PD-1 expression on TILs and an increased PD-L1 expression on TCs (40). Moreover, the application of ICIs in resected NSCLC patients has also been proved an efficient approach by clinical trials and was promising to decrease postoperative recurrence risks (41,42). However, these previous studies primarily focused on the PD-1/PD-L1 pathway, and a robust biomarker for multiple immune checkpoints in NSCLC was still absent. Thus, the IC-Score, Res-Score, and integrated score showed a great potential to screen out the resected NSCLC patients whose

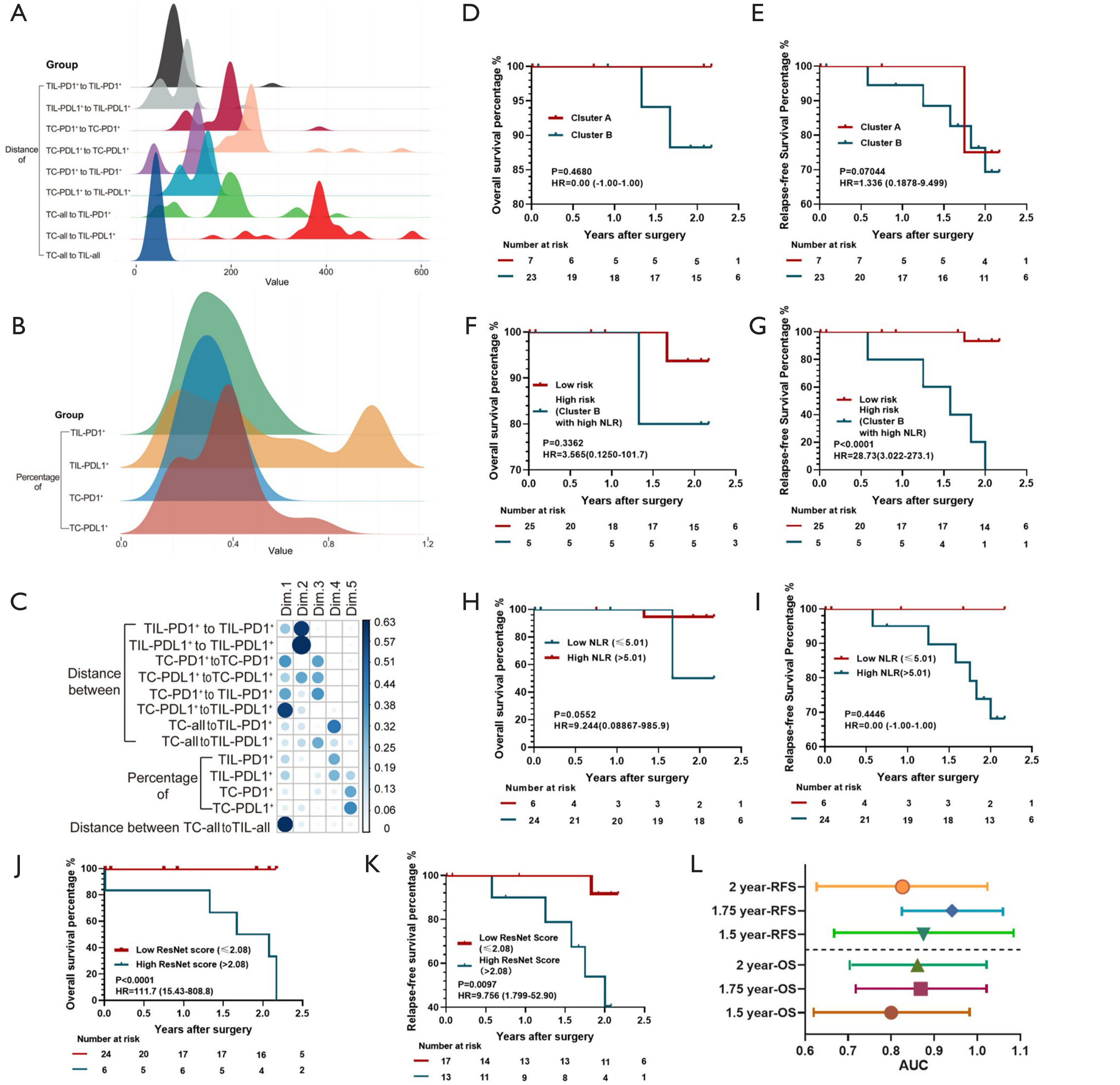


Figure 8 The quantitative and spatial analysis of PD-1/PD-L1, and the performance of the Res-Score in the external cohort. The density curves of spatial features (A) and quantitative features (B) of PD-1/PD-L1 proteins. The spatial features included the distance of TIL_{PD1+} to TIL_{PD1+} (mean ± SEM: 84.77±7.32 px), TIL_{DL1+} to TIL_{PDL1+} (89.76±7.26 px), TC_{PD1+} to TC_{PD1+} (181.7±10.00 px), TC_{PDL1+} to TC_{PDL1+} (248.7±15.14 px), TC_{PD1+} to TIL_{PDL1+} (104.2±7.593 px), TC_{PDL1+} to TIL_{PDL1+} (130.0±5.436 px), TC_{all} to TIL_{PD1+} (188.8±16.41 px), TC_{all} to TIL_{PDL1+} (383.8±15.59 px), and TC_{all} to TIL (40.96±1.202 px). The quantitative features included the ratio of TIL-PD1$^+$ (0.2207±0.02319), TIL-PDL1$^+$ (0.4103±0.06764), TC-PD1$^+$ (0.1747±0.01837), and TC-PDL1$^+$ (0.2370±0.03319). (C) The cos of each feature of PD-1/PD-L1 in each dimension of the PCA. The Kaplan-Meier curves of the PD-1/PD-L1 signature clustering for OS (D) and RFS (E), the combination of NLR and the PD-1/PD-L1 signature for OS (F) and RFS (G), NLR for OS (H) and RFS (I), and the Res-Score for OS (J) and RFS (K). (L) The AUC of the Res-Score for 1.5-year, 1.75-year, 2-year OS and RFS. The lines represent the 95% CI of each AUC value. *The cut-off points of (A, B, G, H) were determined by X-Tile software. PD-1, programmed cell death protein 1; PD-L1, programmed death-ligand 1; Res-Score, ResNet-Score; OS, overall survival; RFS, relapse-free survival; TC, tumor cell; TIL, tumor-infiltrating lymphocyte; SEM, standard error of the mean; HR, hazard ratio; px, pixel; AUC, area under the receiver operating characteristic curve.

risk of relapse and death was relatively high and might benefit from the ICIs targeting the ten immune checkpoints included in this study.

Secondly, the quantitative and spatial analysis of immune checkpoints revealed a clue for future combinational ICIs administration. We found a correlation between the quantity or distribution of OX40/OX40L, LAG-3, and KIR2D. Although there have not been any published research confirming an improved clinical efficacy of the combination of anti-OX40 with anti-LAG3 or anti-KIR2D on NSCLC patients, a series of experiments have posed evidence anti-OX40 could improve T cell differentiation and cytolytic function (43). The relevance between the expression of LAG-3 and KIR2D seemed to be a novel perspective for combinational immunotherapy since current clinical trials used to combine anti-KIR2D with anti-PD1. The function of LAG-3 on NK cells was not well investigated and controversial. However, a factual finding was that inhibition of LAG-3 could increase the secretion of interferon-gamma (IFN-γ), tumor necrosis factor-alpha (TNF-α), macrophage inflammatory proteins-1 alpha (MIP-1α), MIP-1β, and granulocyte-macrophage colony-stimulating factor (GM-CSF) (44). In this way, the expression of LAG-3 could collaborate with the inhibition of NK cell functions from KIR2D.

Thirdly, the quantitative and spatial features of PD-1/PD-L1 proteins in both internal and external cohorts did not play a significant role in the prognostic prediction, although the PD-1/PD-L1 axis was the most targeted protein in immunotherapy. Meanwhile, the inflammation response, such as the status of neutrophils and lymphocytes, is considered as a critical factor in cancer initiation, treatment, and prognosis. Thus, the NLR could assist the PD-1/PD-L1 in predicting OS and RFS, consistent with recent studies (45-47). Whether the combination of preoperative NLR and PD-1/PD-L1 signature correlated with the response to ICIs remains to be investigated, but it still provides hope for discovering therapeutic biomarkers of ICIs.

There remained several limitations in this study. First, the internal cohort of this study included a proportion of missing data. Although we performed multiple imputations to compensate for the missing features, the statistical results might partially differ from the raw data. Secondly, the staining proteins of external IHC slides were less than that of internal IHC slides, which is not conducive to validating the IC-Score in the external testing cohort. Further, the segmentation of TCs and TILs was based on regional labeling instead of individual cells, which needs to be further optimized.

In conclusion, we provided an economical and convenient approach to analyzing the single-plex chromogenic IHC of multiple immune checkpoints, promising to risk-stratify relapse and death in resected NSCLC.

Acknowledgments

Funding: This study was supported in part by a grant of National Natural Science Foundation of China (81802255), Clinical research project of Shanghai Pulmonary Hospital (FKLY20010), Young Talents in Shanghai (2019 QNBJ), 'Dream Tutor' Outstanding Young Talents Program (fkyq1901), Clinical research project of Shanghai Pulmonary Hospital (FKLY20001), Respiratory medicine, a key clinical specialty construction project in Shanghai, Promotion and application of multidisciplinary collaboration system for pulmonary non infectious diseases, Clinical Research Project of Shanghai Pulmonary Hospital (fk18005), Key Discipline in 2019 (oncology), Project of Shanghai Municipal Science and Technology Commission (Project of Municipal Science and Technology Commission), Scientific research project of Shanghai Pulmonary Hospital (fkcx1903), Shanghai Municipal Commission of Health and Family Planning (2017YQ050), Innovation Training Project of SITP of Tongji University, and key projects of leading talent (19411950300). Youth project of hospital management research fund of Shanghai Hospital Association (Q1902037). Figure 1 was created with BioRender.com.

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at http://dx.doi.org/10.21037/tlcr-21-96

Data Sharing Statement: Available at http://dx.doi.org/10.21037/tlcr-21-96

Peer Review File: Available at http://dx.doi.org/10.21037/tlcr-21-96

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/tlcr-21-96). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). It was approved by Shanghai Pulmonary Hospital Ethics Committee (approval number: 15-235), and the written informed consent was obtained from all patients.

References

1. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2007. CA Cancer J Clin 2007;57:43-66.
2. Paz-Ares L, Horn L, Borghaei H, et al. Phase III, randomized trial (CheckMate 057) of nivolumab (NIVO) versus docetaxel (DOC) in advanced non-squamous cell (non-SQ) non-small cell lung cancer (NSCLC). J Clin Oncol 2015;33:LBA109.
3. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus Docetaxel in Advanced Squamous-Cell Non–Small-Cell Lung Cancer. N Engl J Med 2015;373:123-35.
4. Hellmann MD, Nathanson T, Rizvi H, et al. Genomic Features of Response to Combination Immunotherapy in Patients with Advanced Non-Small-Cell Lung Cancer. Cancer Cell 2018;33:843-52.e4.
5. de Miguel M, Calvo E. Clinical Challenges of Immune Checkpoint Inhibitors. Cancer Cell 2020;38:326-33.
6. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. Nat Rev Cancer 2019;19:133-50.
7. Reynders K, De Ruysscher D. Tumor infiltrating lymphocytes in lung cancer: a new prognostic parameter. J Thorac Dis 2016;8:E833-5.
8. Geng Y, Shao Y, He W, et al. Prognostic Role of Tumor-Infiltrating Lymphocytes in Lung Cancer: a Meta-Analysis. Cell Physiol Biochem 2015;37:1560-71.
9. Chand P, Garg A, Singla V, et al. Evaluation of Immunohistochemical Profile of Breast Cancer for Prognostics and Therapeutic Use. Niger J Surg 2018;24:100-6.
10. Gao J, Ren Y, Guo H, et al. A new method for predicting survival in stage I non-small cell lung cancer patients: nomogram based on macrophage immunoscore, TNM stage and lymphocyte-to-monocyte ratio. Ann Transl Med 2020;8:470.
11. Meng J, Zhang J, Xiu Y, et al. Prognostic value of an immunohistochemical signature in patients with esophageal squamous cell carcinoma undergoing radical esophagectomy. Mol Oncol 2018;12:196-207.
12. Massi D, Rulli E, Cossa M, et al. The density and spatial tissue distribution of CD8+ and CD163+ immune cells predict response and outcome in melanoma patients receiving MAPK inhibitors. J Immunother Cancer 2019;7:308.
13. Gruosso T, Gigoux M, Manem VSK, et al. Spatially distinct tumor immune microenvironments stratify triple-negative breast cancers. J Clin Invest 2019;129:1785-800.
14. Giraldo NA, Nguyen P, Engle EL, et al. Multidimensional, quantitative assessment of PD-1/PD-L1 expression in patients with Merkel cell carcinoma and association with response to pembrolizumab. J Immunother Cancer 2018;6:99.
15. Zheng X, Weigert A, Reu S, et al. Spatial Density and Distribution of Tumor-Associated Macrophages Predict Survival in Non-Small-Cell Lung Carcinoma. Cancer Res 2020:canres.0069.2020.
16. Mezheyeuski A, Bergsland CH, Backman M, et al. Multispectral imaging for quantitative and compartment-specific immune infiltrates reveals distinct immune profiles that classify lung cancer patients. J Pathol 2018;244:421-31.
17. Tan WCC, Nerurkar SN, Cai HY, et al. Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. Cancer Commun (Lond) 2020;40:135-53.
18. Dixon AR, Bathany C, Tsuei M, et al. Recent developments in multiplexing techniques for immunohistochemistry. Expert Rev Mol Diagn 2015;15:1171-86.
19. Stack EC, Wang C, Roman KA, et al. Multiplexed immunohistochemistry, imaging, and quantitation: A review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. Methods 2014;70:46-58.
20. Serag A, Ion-Margineanu A, Qureshi H, et al. Translational AI and Deep Learning in Diagnostic Pathology. Front Med (Lausanne) 2019;6:185.
21. Kim MS, Park HY, Kho BG, et al. Artificial intelligence and lung cancer treatment decision: agreement with recommendation of multidisciplinary tumor board. Transl

Lung Cancer Res 2020;9:507-14.
22. Sakamoto T, Furukawa T, Lami K, et al. A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. Transl Lung Cancer Res 2020;9:2255-76.
23. Wang X, Li Q, Cai J, et al. Predicting the invasiveness of lung adenocarcinomas appearing as ground-glass nodule on CT scan using multi-task learning and deep radiomics. Transl Lung Cancer Res 2020;9:1397-406.
24. Zhang X, He Y, Jia K, et al. Does selected immunological panel possess the value of predicting the prognosis of early-stage resectable non-small cell lung cancer? Transl Lung Cancer Res 2019;8:559-74.
25. He Y, Rozeboom L, Rivard CJ, et al. PD-1, PD-L1 Protein Expression in Non-Small Cell Lung Cancer and Their Relationship with Tumor-Infiltrating Lymphocytes. Med Sci Monit 2017;23:1208-16.
26. He Y, Rozeboom L, Rivard CJ, et al. MHC class II expression in lung cancer. Lung Cancer 2017;112:75-80.
27. He Y, Jia K, Dziadziuszko R, et al. Galectin-9 in non-small cell lung cancer. Lung Cancer 2019;136:80-5.
28. He Y, Bunn PA, Zhou C, et al. KIR 2D (L1, L3, L4, S4) and KIR 3DL1 protein expression in non-small cell lung cancer. Oncotarget 2016;7:82104-11.
29. He Y, Zhang X, Jia K, et al. OX40 and OX40L protein expression of tumor infiltrating lymphocytes in non-small cell lung cancer and its role in clinical outcome and relationships with other immune biomarkers. Transl Lung Cancer Res 2019;8:352-66.
30. He Y, Yu H, Rozeboom L, et al. LAG-3 Protein Expression in Non&#x2013;Small Cell Lung Cancer and Its Relationship with PD-1/PD-L1 and Tumor-Infiltrating Lymphocytes. J Thorac Oncol 2017;12:814-23.
31. Jia K, He Y, Dziadziuszko R, et al. T cell immunoglobulin and mucin-domain containing-3 in non-small cell lung cancer. Transl Lung Cancer Res 2019;8:895-906.
32. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.
33. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:190511946 2019.
34. Baheti B, Innani S, Gajre S, et al., editors. Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 14-19 June 2020.
35. Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:160308029 2016.
36. Li P, Stuart EA, Allison DB. Multiple Imputation: A Flexible Tool for Handling Missing Data. JAMA 2015;314:1966-7.
37. Newgard CD, Lewis RJ. Missing Data: How to Best Account for What Is Not Known. JAMA 2015;314:940-1.
38. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res 2004;10:7252-9.
39. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16:385-95.
40. Markowitz GJ, Havel LS, Crowley MJ, et al. Immune reprogramming via PD-1 inhibition enhances early-stage lung cancer survival. JCI Insight 2018;3:e96836.
41. Forde PM, Chaft JE, Smith KN, et al. Neoadjuvant PD-1 Blockade in Resectable Lung Cancer. N Engl J Med 2018;378:1976-86.
42. Kwiatkowski DJ, Rusch VW, Chaft JE, et al. Neoadjuvant atezolizumab in resectable non-small cell lung cancer (NSCLC): Interim analysis and biomarker data from a multicenter study (LCMC3). J Clin Oncol 2019;37:8503.
43. Linch SN, McNamara MJ, Redmond WL. OX40 Agonists and Combination Immunotherapy: Putting the Pedal to the Metal. Front Oncol 2015;5:34.
44. Narayanan S, Ahl PJ, Bijin VA, et al. LAG3 is a Central Regulator of NK Cell Cytokine Production. bioRxiv 2020:2020.01.31.928200.
45. Wang X, Cao L, Li S, et al. Combination of PD-L1 expression and NLR as prognostic marker in patients with surgically resected non-small cell lung cancer. J Cancer 2019;10:6703-10.
46. Tashima Y, Kuwata T, Yoneda K, et al. Prognostic impact of PD-L1 expression in correlation with neutrophil-to-lymphocyte ratio in squamous cell carcinoma of the lung. Sci Rep 2020;10:1243.
47. Banna GL, Signorelli D, Metro G, et al. Neutrophil-to-lymphocyte ratio in combination with PD-L1 or lactate dehydrogenase as biomarkers for high PD-L1 non-small cell lung cancer treated with first-line pembrolizumab. Transl Lung Cancer Res 2020;9:1533.

Cite this article as: Guo H, Diao L, Zhou X, Chen JN, Zhou Y, Fang Q, He Y, Dziadziuszko R, Zhou C, Hirsch FR. Artificial intelligence-based analysis for immunohistochemistry staining of immune checkpoints to predict resected non-small cell lung cancer survival and relapse. Transl Lung Cancer Res 2021;10(6):2452-2474. doi: 10.21037/tlcr-21-96