Pretreatment radiomic biomarker for immunotherapy responder prediction in stage IB-IV NSCLC (LCDigital-IO Study): A multi-center retrospective study

Appendix S1: The Detailed Inclusion and Exclusion Criteria.

For cohort D1-D3, the inclusion criteria were as follows: (1) more than 18 years old, (2) first dose of anti-PD- (L)-1 immunotherapy were received in Guangdong Provincial People's Hospital or Jiangxi Cancer Hospital, (3) Patients with advanced NSCLC confirmed pathologically, (4) a follow-up period of at least 6 months. The exclusion criteria were as follows: (1) lack of chest thin-slice (≤5mm) enhanced CT image within 30 days before the first immunotherapy, (2) missing or incomplete necessary clinicopathological variables, (3) poor image quality (e.g., motion and speed propagation, refraction artifacts or low-resolution image), (4) patients with unilateral multifocal cancers, (5) received other therapy meanwhile, (6) suffering other malignancies simultaneously, (7) pulmonary lesions were poorly discriminated from other lesions or adjacent tissues.

Appendix S2: Summary for Package Usage.

For the radiomic signature extraction, we used "SlicerRadiomics" package on the platform of 3D Slicer (http://www. slicer. org). The Least absolute shrinkage and selection operator algorithm was adopted using R package named "glmne t" (version 4.1.3; http://www.r-project.org/). The AUC is calculated by the R package named "pROC". The decision curve was calculated via the R package named "decisionCurve". We drew Kaplan-Meier curves by the R packages named "survival" and "survminer". For the multivariate analysis of clinical characteristics and the radiomic biomarker CRS, we determined the interaction of the variables in survival analysis using SPSS version 23.0 (IBM,

Armonk, NY). The R package "limma" was used to identify differentially expressed genes (DEGs) between high-risk and low-risk groups based on the radiomic signature. Single sample Gene Set Enrichment Analysis (ssGSEA) was conducted via R package “GSVA”.

Appendix S3: Acquisition of Computed Tomography (CT) Images.

Thoracic CT examinations in GDPH were performed using one of two multidetector CT systems (LightSpeed 8; GE Healthcare, Hino, Japan; 64-slice VCT, GE Healthcare, Milwaukee, WI). Scanning parameters were as follows: 120 kVp, 150–200 mA, detector collimation of 8 × 1.25 mm or 64 × 0.625 mm; field of view of 350 mm × 350 mm; matrix of 512 × 512, and reconstruction thickness of 1.25-mm with a 1.25-mm reconstruction interval. Nonionic iodinated contrast material (370 mg iodine/mL, Ultravist; Bayer Pharma, Berlin, Germany) was injected into the antecubital vein at a dose of 1.3–1.5 mL/kg of body weight at a rate of 2.5 mL/s using an automated injector (Ulrich CT Plus 150; Ulrich Medical, Ulm, Germany). CT scanning was performed with a 25-second delay after contrast agent injection. In JXCH, CT examinations was performed in a single CT system (Somatom Definition Edge, Siemens Healthineers, Germany, 64-slice VCT). The parameters of the images were as follows: 80-130 kV; 130 mAs; rotation time of 0.4-/0.6s; detector collimation of 64×0.625 mm; field of view of 350 mm × 350 mm; matrix, 512 × 512. Gastrografin was applied as contrast material (76% compound meglumine diatrizoate injection, 20 ml/ampule, containing 15.2 g meglumine diatrizoate and sodium diatrizoate; Shanghai Xudong Haipu Pharmaceutical Co., Ltd., Shanghai, China) and was injected into the antecubital vein at a dose of 1.2–1.6 mL/kg of body weight at a rate of 3 mL/s using an automated injector (SinoPower-S; Sinomdt Medical Equipment Co., Ltd., Shenzhen, China). CT scanning was

performed with a 30-second delay after contrast agent injection. The CT images were retrieved using the picture archiving and communication system (PACS; Carestream, Canada) at each institution.

Appendix S4: Imaging Processing and ROI Segmentation.

Before feature extraction, the voxel size of all slices was resampled to 1mm × 1 mm× 1 mm. A total of 851 features were extracted from each mask via SlicerRadiomics package on the 3DSlicer platform (version 4.10.2, www.slicer.org), including 14 shape-based features, 18 first-order statistical features, 75 textural features from original images and 744 derived features. The features were then normalized with the z-scores of the training and validation cohorts using the mean and standard deviation derived from the features of the training cohort. The formula used is as follows: $Z=(X-\bar{x})/SD$ where X represents the value of each selected feature in a patient, while Z represents the corresponding normalized value. $\bar{x}$ represent the mean of the values of the feature and SD is the standard deviation in the training cohort. The segmentation was implemented by a preliminary semi-automatic segmentation module according to the level tracing and intensity threshold segmentation tool. Then the manual corrections such as relabeling and hole filling were done by S.Y.W (with 14 years of experience) and G.Y.W (with 18 years of experience). Radiologist L.L. (with 23 years of experience) assessed all tumor segmentations. Any disagreements were resolved by discussions between the three radiologists mentioned and three thoracic surgeons (Q.G.B., Z.D.K., and H.Y.Z.). The modules of logical operation and hollow helps us to automatically generalize the peri-tumoral segmentations according to the segmented tumoral region.

Appendix S5: Feature extraction and selection.

There are 4255 quantitative radiomic features in all, which include first-order statistics, shape, gray-level cooccurrence matrix (GLCM), gray-level size zone matrix (GLSZM), gray-level dependence matrix (GLDM), neighborhood gray-tone difference matrix (NGTDM) and the corresponding wavelet-transformed features, which were extracted from five segmented regions (tumoral/peritumoral 0-5mm/5-10mm/10-15mm/15-20mm, 851 features for each segment). RIDER dataset is an online available dataset[1], which includes 31 patients, each patient having been scanned twice on a CT scanner with an interval of approximately 15 minutes. For each radiomic feature extracted, the intraclass correlation coefficient (ICC) was calculated to quantify reproducibility between the test–retest scans (RIDER). Because ICC describes the similarity of units in the same group, features with high ICC values are thus more reproducible and potentially more robust to variations in CT scanners and acquisition parameters. Only the features with ICC cutoff >0.75 were identified as stable and reproducible features (n=3630). In order to assess the multi-collinearity of the characteristics, the Spearman correlation analysis was conducted and the correlation coefficients among the features were determined. Only the feature with a superior diagnostic performance was kept if any pair of features had a coefficient value of > 0.80 or < -0.80. Within all stable features (n=2285) we continued the further feature selection. The least absolute shrinkage and selection operator (LASSO) algorithm was next applied to filter out redundant and unpredictive features. We used step-wise binominal logistic regression for the selected features to construct the radiomic signatures.

Appendix S6: The Effect of Tumor Volume on CRS and Clinical Outcome.

To evaluate the association between radiomic score and tumor volume, we calculated the

Pearson coefficient between them (seen in **Supple table.6a)**, and no significant association was observed (p>0.05). Tumor volume is calculated by $V = \pi(\mathrm{a} * \mathrm{b} * \mathrm{c})/6,$ where a, b, c are the linear dimensions of the tumor[2]. To assess the confounding effect of tumor volume on CRS for DCB prediction, a binary logistics regression was conducted where tumor volume was set as a control variable. From **supplemental table.6b,** we found tumor volume is not associated with patients' response status in neither univariate nor bivariate analysis. For long-term survival, a univariate analysis was conducted for the tumor volume on PFS and OS. The analysis showed poor stratification for both PFS (p=0.948) and OS (p=0.753). Subsequently we combined tumor volume with CRS for a two-variate analysis on OS and PFS to evaluate the confounding effect. We found that tumor volume has nearly no effect on long-term survival (PFS: HR=1, p=0.997; OS: HR=1, p=0.753, **Supple table.6c**) while CRS still shows good stratification for both PFS (HR=0.487, p=0.002) and OS (HR=0.484, p=0.022).

Appendix S7: Assessment of Model and Statistics Methods.

1. The Hosmer-Lemeshow test was used to assess model goodness of fit, where we compared the computed Hosmer-Lemeshow statistic to a chi-squared distribution with 8 degrees of freedom to calculate the p-value. Hosmer-Lemeshow statistic was calculated as follows:

$$H = \sum_{q=1}^{10} \left(\frac{(Observed.DCB - Expected.DCB)^2}{Expected.DCB} + \frac{(Observed.\ not\ DCB - Expected.not\ DCB)^2}{Expected.not\ DCB} \right)$$

2. DCAs was applied to assess the utility of the three signatures by calculating the net benefit at various probability thresholds[3]. In our study, threshold probability is the value

of the lowest suspected probability of DCB where the physician still may advise the patient to receive immunotherapy after balancing the benefit and harmfulness. As illustrated in the background section, the immunotherapy response rate is still limited. And the adverse effect remains a major concern with 28.7% of 3-5 grades treatment related adverse events for patients receiving ICIs monotherapy[4]. Therefore, the probability threshold in clinical practice would not be low (for example, 10%), and it varies among different physicians and patients when they decide to receive immunotherapy. According to the decision curves shown in Figure 3D and Figure S9, the CRS and RADCli signatures both always showed more net benefit than the clinical signature in predicting the probability of DCB. And they are consistently applicable when the probability threshold was between 5% to 97%, which was wide-range and consistent with clinical practice.

3. Lesion Choice Strategy Subgroup Analysis

We applied the thoracic CT image for radiomic analysis, which was routinely and widely used in clinical practice. The pulmonary lesion choice strategy is certain and adaptive to the clinical practice, that is: First, only pulmonary lesion was chosen. Secondly, principally the primary pulmonary lesion was chosen as target lesion. Thirdly, the largest pulmonary lesion was the candidate choice for patients whose primary lesion cannot be certainly determined since first admission. To evaluate the impact on prediction brought by lesion choice strategy, we performed a subgroup analysis for patients of different lesion choices. Two choice strategies do not show significant difference on DCB prediction in both cohort D2 ($AUC_{primary}$=0.629, 95% CI: 0.435-0.824; $AUC_{largest}$=0.767,

95% CI: 0.460-1, p=0.466) and D3 ($AUC_{primary}$=0.887, 95% CI: 0.772-1, $AUC_{largest}$=0.714, 95% CI: 0.379-1, p=0.357, Fig. S10). Due to the unbalanced ratio of patients of two lesion choice strategies, we combined two independent test sets D2 and D3 to make the two groups more comparable (N=105, $N_{primary}$ = 81, AUC=0.769, 95% CI: 0.654-0.883; $N_{largest}$=24, AUC=0.736, 95% CI: 0.523-0.948). The DeLong test showed that no significant difference of the prediction was observed between the two groups (p= 0.791).

4. Survival analysis grouping

   The predicted DCB and NCB groups were based on the optimal cut-off of this score on the training cohort. The optimal cut-off is obtained by maximizing the Youden index (Youden index = Sensitivity + Specificity − 1), ensuring the best overall performance of the signature. Based on the prediction, two predicted groups were generated and showed distinct survival outcomes in survival analysis.

5. The two-way random ICC test was applied to measure the reliability of the radiomic features between two-observer image segmentation and feature extraction process. The agreement levels are defined regarding ICC values: excellent (ICC≥0.81), good (0.61<ICC<0.8), moderate (0.41<ICC<0.60), and poor (ICC≤0.40). TableS8 summarizes the results of the inter-rater agreement analysis. Radiomic features in the CRS show good to excellent inter-rater reliability (ICC=0.72 to 0.99).

Appendix 8: Algorithm Selection.

Different algorithms have been tested during pre-study period. We conducted a pre-experiment comparing four different algorithms configurations (seen in Table S7). In this analysis, we selected the final configuration by comparing their performance on the validation

set. We chose LASSO and stepwise logistic regression as candidate classifier because this configuration reached the highest performance with the least amount of overfitting.

Captions for supplemental tables and figures.

Table S1. Clinicopathological characteristics of D1-D3 cohorts

Table S2. The performance of different radiomic signatures in cohorts

Table S3. The comparison of performance of different peri-tumoral radiomic signatures based on AIC and DeLong test

Table S4. The coefficient of p value and Hazard ratio for Log-Rank test in clinical variables on PFS and OS

Table S5. Odds ratio of clinical variables via univariate and multivariate logistics analysis

Table S6. Evaluation of tumor volume on CRS and clinical outcome

Table. S7. The performance of different algorithms on training set and validation set.

Table. S8. Table S8 Inter-observer variability analysis of radiomic feature extraction.

Figure S1. Study design. The discovery dataset was used to train the radiomic signature. Four cohorts were used for validation. The Cancer Genome Atlas dataset comprised RNA-seq data and the corresponding imaging data from The Cancer Imaging Archive. The surgical cohort contained patients receiving radical surgery after receiving immunotherapy. RNA-seq=RNA sequencing. DICOMS=Digital Imaging and Communications in Medicine.

Figure S2. LASSO coefficient profiles of the retained 851 features. A coefficient profile plot was produced against the log (λ) sequence. Vertical line was drawn at the value selected using 10-fold cross-validation, where optimal λ resulted in 2 tumoral-features and 7 peritumor features with nonzero coefficients.

Figure S3. Calibration curve analysis for CRS in training and validation cohorts.

Figure S4. Performance and evaluation of clinical signature. (A) ROC curves with AUC of 0.76 (95%CI: 0.674-0.846) in training cohort and (B) 0.66 (95%CI: 0.509-0.811) in validation cohort. (C) Calibration curve for training and (D) validation cohort.

Figure S5. Calibration curve of RADCli(nomogram) in (A) training cohort and (B) validation cohort.

Figure S6. Kaplan-Meier survival curves stratified by optimal cutoff of PD-L1 expression for OS(A) and PFS(B) in PD-L1 subset, Kaplan-Meier survival curves stratified by classification based on CRS for OS(C) and PFS(D) for the PD-L1 subset (n=138).

Figure S7. The Kaplan-Meier survival analysis for CRS in TCGA-NSCLC cohort.

Figure S8. Gene Ontology annotation of differentially expressed genes in TCGA-NSCLC cohort.

Figure S9. Decision curve analysis for the RADCli nomogram (red), CRS signature (blue), and clinical model (green) with corresponding confidence intervals represented by dashed lines.

Figure S10. The ROC analysis of subgroup evaluation for different lesion choice strategies in external cohorts. (A) Cohort D2. (B) Cohort D3. (C) Cohort D2 and D3.

Reference.

1 Armato SG, Meyer CR, McNitt-Gray MF, *et al.* The Reference Image Database to Evaluate Response to therapy in lung cancer (RIDER) project: A resource for the development of change-analysis software. Clin Pharmacol Ther. 2008; **84**: 448–56.

2 Mayr NA, Taoka T, Yuh WTC, *et al.* Method and timing of tumor volume measurement

for outcome prediction in cervical cancer using magnetic resonance imaging. *Int J Radiat Oncol Biol Phys* 2002; **52**: 14–22.

3 Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak* 2008; **8**. DOI:10.1186/1472-6947-8-53.

4 Alifu M, Tao M, Chen X, Chen J, Tang K, Tang Y. Checkpoint inhibitors as dual immunotherapy in advanced non-small cell lung cancer: a meta-analysis. *Front Oncol* 2023; **13**: 1146905.