Peer Review File

Article information: https://dx.doi.org/10.21037/tlcr-24-7

Reviewer A

The authors submitted a study that employed delta-features to establish the prognosis of patients with NSCLC undergoing immunotherapy. The paper is neat and well-presented. The study is very interesting and well-performed; the Methods seem robust and adequate for a study based on radiomics.
I have a few minor points:
Comment 1: - The patients from the second Institution employed to assess the performance of the model are not the "validation dataset" but the "test dataset". The validation set is a partition of the original cohort from the same Institution employed to assess the performance before testing on an independent set.

Reply 1: We highly appreciate the Reviewer's kind reminder. We have switched the term "validation dataset" with "test dataset" throughout the revised manuscript, figures, and tables.

Comment 2:- In the Introduction, the Authors should mention the response criteria designed specifically for immunotherapy such iRECIST and irRECIST.

Reply 2: We highly appreciate the Reviewer's constructive comments. We have mentioned the iRECIST criteria designed specifically for immunotherapy in the revised manuscript. In addition, we have reevaluated the prediction performance based on iRECIST criteria for anti-PD-1 therapy in the revised manuscript.
Changes in the text:
The immune Response Evaluation Criteria in Solid Tumors (iRECIST) represents an enhanced and revised iteration of the RECIST guidelines. Its widespread adoption in clinical trials worldwide is owed to its improved and standardized assessment of tumor response, specifically incorporating immune-based treatment responses like hyperprogression and pseudoprogression (9, 10). Nonetheless, the iRECIST criteria overlook changes in various tumor characteristics beyond size, such as tumor viability, metabolic activity, and tumor density, which could be pertinent to tumor response (11, 12). Hence, there is an urgent need for an alternative approach to anticipate response to anti-PD-1 therapy ICIs.

Comment 3: - I do not understand how the Authors employed Person's correlation analysis. Could you be clearer?

Reply 3: We sincerely appreciate the comments from the Reviewer. The purpose of Person's correlation analysis in this study is to eliminate the radiomics features with

high correlation(r>0.8) to exclude redundant variables. As far as we know, Pearson's correlation has been widely used in the variable selection of radiomics research (such as the research: Radiomics analysis for predicting the prognosis of colorectal cancer from preoperative 18F-FDG PET/CT. J Transl Med. 2022 Feb 2;20(1):66; Integrating tumor and nodal radiomics to predict lymph node metastasis in gastric cancer. Radiother Oncol. 2020 Sep;150: 89-96). We have included explanations and cited the references in the methods section.
Changes in the text:
Pearson's correlation analysis was used to eliminate the radiomics features with high correlation(r>0.8), thereby excluding redundant variables (25).

Comment 4: - Line 307: "confirmed by two independent dataset". I think that the Authors cannot claim this. The Authors used a dataset for training and a dataset for assessing generalizability.

Reply 4: We highly appreciate the Reviewer's constructive comments. We have changed the sentences as *"Moreover, these findings were further confirmed by an external independent database."* the revised manuscript.

Comment 5: - Lines 314-315: "T1-weighted magnetic resonance (MR) sequences was associated with immunophenotyping". This sentence is a bit vague, I suggest re-phrasing it.

Reply 5: We highly appreciate your critique. We've reformulated the sentence to make it clearer.
Changes in the text: *A recent study (28) demonstrated the firstorder_Median feature extracted on enhanced T1-weighted MR sequences has been demonstrated to correlate with immunophenotyping and serves as a radiomics biomarker for predicting overall survival (OS) in patients diagnosed with intrahepatic cholangiocarcinoma.*

Comment 6: - Lines 343-344: I do not agree with the Authors about the fact that contrast media might obscure imaging features that reflect potential tumor heterogeneity. In immunotherapy the response is immune-mediated, and it depends on inflammation which can be better seen using contrast media. Please revise.

Reply 6: We highly appreciate the Reviewer's constructive comments. We have deleted the related statement.

Reviewer B

The paper presented an interesting study. One important aspect of the study is that they used an independent dataset to test the performances of the models to report how

well the models can be generalized to new data. If successful, research along this line can play an important role in providing personalized treatment leveraging knowledge and information beyond conventional clinical decision-making, especially if more data from more centers can be combined to learn models on a larger scale. The study was well-designed. The data analysis was comprehensive. Some specific comments are given in the following.
Comment 1: Line 57: Define "ICI" before using it for the first time

Reply 1: We highly appreciate the Reviewer's kind reminder. We have defined "ICI" as using it for the first time.

Comment 2: Line 58: How were the cutoff values of "ICC<0.8" and "r>0.8" decided? Why not 0.7 or 0.9?

Reply 2: We chose an ICC< 0.8 as the features of low reliability and an r >0.8 as redundant variables, which were based on the references (Pleil JD, Wallace MAG, Stiegel MA, Funk WE. Human biomarker interpretation: the importance of intraclass correlation coefficients (ICC) and their calculations based on mixed models, ANOVA, and variance estimates. J Toxicol Environ Health B Crit Rev. 2018;21(3):161-180. Radiomics analysis for predicting the prognosis of colorectal cancer from preoperative 18F-FDG PET/CT. J Transl Med. 2022 Feb 2;20(1):66), which has been cited in the revised manuscript.

Comment 3: Line 179: Data were collected from different machines. I wonder if the authors or any previously published studies have compared the influence of different machines on image quality or accuracy.

Reply 3: We highly appreciate the Reviewer's comments. To the best of our understanding, while studies have explored the impact of various scanning protocols on radiomics, there remains a notable absence of research concerning the influence of different machines on radiomics quality or accuracy. (Huisman M, Akinci D'Antonoli T. What a Radiologist Needs to Know About Radiomics, Standardization, and Reproducibility. Radiology. 2024 Feb;310(2): e232459.)
In our study, we meticulously standardized scan protocols across different CT machines, aligning parameters such as tube current, tube voltage, and reconstruction algorithms to ensure uniform image quality.

Comment 4: Line 183: Why were different slice thicknesses used? The authors mentioned two reconstructions at different thicknesses, which is confusing. Why were these reconstructions necessary and why were two needed? Was the purpose to be able to obtain 3D radiomic features from 3D volumes? In addition, what reconstruction algorithms were used? Did they cause any smoothing effect on the volume?

Reply 4: We sincerely apologize for the confusion caused.

(1) Due to the retrospective nature of data collection in this study, there was variation in the thickness of reconstructed slices across different CT scanners.

(2) We have meticulously revised the manuscript, removing the inappropriate description. Additionally, we employed resampling with spline interpolation to ensure uniform voxel size for the radiomics features. This clarification has been incorporated into the revised manuscript as follows.

Changes in the text:

Images were reconstructed at a slice thickness of 1.5 or 1 mm with an increment of 1.5 or 1 mm. No contrast medium was used. Then, all images were resampled with a slice thickness of 1.5 mm with the same increment to ensure uniform voxel size(22).

Ligero M, Garcia-Ruiz A, Viaplana C, et al. A CT-based Radiomics Signature Is Associated with Response to Immune Checkpoint Inhibitors in Advanced Solid Tumors. Radiology. 2021;299(1):109-119.

Comment 5: Line 197: It is ok not to list all features. However, what types of features were included should be summarized. For instance, were all features extracted from 3D volume or 2D slices? Were there anatomical shape features such as lengths and areas?

Reply 5: We highly appreciate the Reviewer's constructive comments. We have incorporated the types of features into the methods section, along with specifying that all features were extracted from the 3D volume, as also detailed in the methods section.

Changes in the text:

The heatmap of the 1,037 radiomics feature could be categorized based on image type and feature class. By image type, there were 93 gradient features, 93 LoG features, 107 original features, and 744 wavelet features. By feature class, there were 198 first-order features, 264 GLCM features, 154 GLDM features, 176 GLRLM features, 176 GLSZM features, 55 NGTDM features, and 14 Shape features. All features were extracted from the 3D volume.

Comment 6: Line 200: It looks like the features were from 2D segmented regions. If so, why was reconstruction necessary?

Reply 6: All features were extracted from the 3D volume, as also detailed in the methods section.

Comment 7: Line 229: How were features of different types (for instance clinical features and delta radiomics features) combined so their relative importance in the overall metric is reasonable?

Reply 7: We highly appreciate the Reviewer's constructive comments.
We incorporated the selected radiomics features and clinical features into the Lasso-cox risk regression analysis, thus obtaining a mix model, as previous reported (Ortega C, Eshet Y, Prica A, et al Combination of FDG PET/CT Radiomics and Clinical Parameters for Outcome Prediction in Patients with Hodgkin's Lymphoma. Cancers (Basel). 2023 Mar 30;15(7):2056). This clarification has been incorporated into the revised manuscript as follows.

Changes in the text:

Then, we incorporated the selected radiomics features and clinical features into the Lasso-cox risk regression analysis, thus obtaining a mix model, as previous reported (29).

Comment 8: Line 252: Please describe how normalization was performed on these features.

Reply 8: We highly appreciate the Reviewer's comments. Since clinical features are the classification criteria, normalization was not performed on these features.

Comment 9: Lines 254-256: The training accuracy seems higher than test and validation datasets, which indicates possible overfitting. Did authors consider overfitting? How were the model hyper-parameters such as training stopping criteria determined?

Reply 9: We highly appreciate the reviewer's concerns. Due to the complexity of the survival analysis model and our small sample size, variation in the performance was inevitably seen among the training, test and validation sets for our developed models. However, the C-index of the test and validation sets always fell into the range of the C-index in the fivefold cross-validation of the training set, indicating that our model is generally stable. This point has been explained in the limitation section.

Changes in the text:

Moreover, given the intricate nature of the survival analysis model and our limited sample size, variations in performance were observed across the training, test, and validation sets. Nevertheless, the C-index consistently fell within the range established by fivefold cross-validation on the training set, indicating overall stability of our model.

Comment 10: Lines 262 and 267: Why 6 and 8 features were selected? How are these selected features compared to the identified features used in published studies for similar purposes? Do they fall in the same feature category ?

Reply 10: We highly appreciate the Reviewer's constructive comments.
(1) We selected the lambda corresponding to the maximum C-index via Lasso, followed by feature selection using Cox proportional hazards regression: *The*

remaining 145 radiomics features were selected by LASSO for further assessment. Then, six most contributing radiomic features were selected to establish a pre-treatment radiomics models by cox proportional hazards regression for predicting OS.
(2) We have compared these selected features to the identified features used in published studies for similar purposes in the discussion part: *A recent study (29) demonstrated the firstorder_Median feature extracted on enhanced T1-weighted MR sequences has been demonstrated to correlate with immunophenotyping and serves as a radiomics biomarker for predicting OS in patients diagnosed with intrahepatic cholangiocarcinoma.*

Comment 11: Lines 279-290: Same issue here. What were the criteria to decide how many features to use?

Reply 11: We selected the lambda corresponding to the maximum C-index via Lasso, followed by feature selection using Cox proportional hazards regression: *For PFS, five most contributing radiomic featuresfive delta-radiomics features (delta_wavelet_LHL_firstorder_Median , delta_original_shape_LeastAxisLength , delta_wavelet_LLL_glcm_Contrast , delta_wavelet_HLL_glcm_Idmn , delta_original_glcm_DifferenceAverage) were imported into the final model . by cox proportional hazards regression.*

Comment 12: Figure 1: Font size is too small to read. Authors should consider rearranging the blocks to improve readability.

Reply 12: We highly appreciate the Reviewer's constructive comments. We have adjusted Figure 1 to ensure the resolution.

Comment 13: Figure 4: font size for A and C is too small. The results on the independent validation set seem to be consistently lower than the test data. Do authors have any insights on why that's the case? Is it possible that it has something to do with different data representation from different centers?

Reply 13: We highly appreciate the Reviewer's constructive suggestions. We have amplified font size for A and C.
Furthermore, we fully concur with the Reviewer. This result may attribute to the fact that the CT scanning machines of different manufacturers was used in independent validation datasets compared to with the model building data. We have discussed this point in the revised manuscript.

Changes in the text:
Fourth, the independent test set consistently yields lower results compared to the validation data in our models, likely due to the use of CT scanning machines from different manufacturers at the two centers.