Peer Review File

Article Information: https://dx.doi.org/10.21037/tcr-22-2683

Reply to Reviewer A's Comments:

#1 This is a retrospective observational study that develops a model to delineate radiomic CT imaging traits between somatic EGFR mutation positive and negative cases of lung cancer. Radiomics is an established field which uses features detected computationally that may not be visible or comparable to the human eye. The manuscript is mostly clearly written and has a valid approach. It is a small study that provides data on radiomic and genetic correlations that could be explored further in the future to predict presence of EGFR mutations based on imaging findings. This has relevance as EGFR status influences treatment decisions. The main limitations of the study are its small sample size and single site data set (no external validation) and some technical limitation in terms of the extent of the approaches employed eg no justification of chosen feature reduction/ML approach/comparison of different approaches. There is previous literature on this concept e.g. 10.3389/fonc.2020.00028 (Front oncology), https://doi.org/10.1007/s00330-019-06024-y (oncology), and https://doi.org/10.3389/fonc.2020.54295 (front oncology) so the novelty is limited. It's applicability is also not well defined as ultimately tissue or blood (ctDNA) genetic information would ultimately be required for treatment decisions and obtaining this information is not unreasonably challenging. This field may benefit from combining the data available into a larger collaborative series.

Response: Thank you for your very helpful comments. We acknowledge that this study included the small size of patients and no external validation. As you pointed out, external validation is preferable for evaluating the machine-learning-based prediction model, which corresponds to Types 3 and 4 in the TORIPOD guidelines. However, if available data are limited, certain resampling techniques such as bootstrapping and cross-validation are recommended in the guideline (that is, Type 1b), thus 5-fold cross validation was selected to evaluate the prediction performance in this study. Regarding the selection of ML model, we selected the Random Forest classifier from various ML techniques because we wanted to compare the feature importance between the clinical variables and radiomic features in the Combined model. In general, other ML techniques such as support vector machine [28 in the revised manuscript] and deep learning [26] can be used to predict treatment outcomes, but how much candidate variables contributed to the prediction is unclear. Our study demonstrated that the importance of two radiomic features were higher than those of the relevant clinical features (gender and smoking history) in the combined model, suggesting that the radiomic features have a higher impact on detecting EGFR mutations over the clinical features. Further, Boruta algorithm was chosen for feature selection because it is commonly used for constructing the bagging-based prediction model [#1-

3].

Thank you for letting us know that there have been similar reports in the past. Even in these reports, only few of them comprised the surgical population (mostly Stage III or IV were included in those report), thereby limiting the results for early-stage population. Our study focused on the preoperative CT images from early-stage NSCLC (Stage I or II), aiming for potential usage in neoadjuvant settings (e.g., NeoADAURA: NCT04351555).

We agree with the fact that tissue or blood (ctDNA) genetic information would ultimately be required for treatment decisions. In recent years, ctDNA technology has advanced dramatically, but there is the possibility of false-negative result from the liquid biopsy in NSCLC [33], thereby considering the tissue-based analysis for EGFR mutations as the reference standard [34]. In this point, radiomics-based prediction of EGFR mutations could help the choice of preoperative treatment in neoadjuvant setting like ctDNA, if the quantity or quality of biopsy samples is not enough for DNA testing.

Some descriptions and references have been added to the revised manuscript as follows:

Page 14, Line 253; DISCUSSION

"Although we performed cross-validation according to TRIPOD guidelines (22), our findings need further validation with a large sample set."

Page 14, Line 271; DISCUSSION

"Finally, we did not consider other machine-learning techniques other than Random Forest for prediction. Jia TY et al. used Random Forest classifier to predict EGFR mutations using radiomic features and clinical variables (29), and demonstrated that relatively higher predictive performance was observed (AUC = 0.828), which is consistent with our results (AUC = 0.83)."

Page 12, Line 222; DISCUSSION

Previous studies have shown the possibility of EGFR mutation status in NSCLC utilizing CT images.(24–29)

Please note that the reference (https://doi.org/10.3389/fonc.2020.54295; front oncology) you cited did not appear in web site, thus we could not include it in the References.

Page 13, Line 231; DISCUSSION

"Moreover, even in the liquid biopsy (ctDNA), there is the possibility of false-negative result in NSCLC (33), thereby considering the tissue-based analysis for EGFR mutations as the reference standard (34)."

References

#1 Shen X, Yang F, Yang P, et al. A Contrast-Enhanced Computed Tomography Based Radiomics Approach

for Preoperative Differentiation of Pancreatic Cystic Neoplasm Subtypes: A Feasibility Study. Front Oncol. 2020;10:248.

#2 Wang S, Sun Y, Mao N, et al. Incorporating the clinical and radiomics features of contrast-enhanced mammography to classify breast lesions: a retrospective study. Quant Imaging Med Surg. 2021;11(10):4418-4430.

#3 Gangil T, Shahabuddin AB, Dinesh Rao B, et al. Predicting clinical outcomes of radiotherapy for head and neck squamous cell carcinoma patients using machine learning algorithms. J Big Data. 2022;9:25.

#2 Please clarify term c-stage, in the other settings this might be termed T-stage

Response: Thanks. The term of "c-stage" has been changed to "clinical stage" in the revised manuscript and Table 1.

#3 Line 71 - In small peripheral tumours, other biopsy approaches such as CT guided biopsy would be more typical. A statement noting this feels to be missing if the authors are stating bronchoscopic biopsy limitations.

Response: Thank you for your comment. We agree with your opinion, thus the description regarding the CT-guided biopsy has been added to the revised manuscript as follows:

Page 5, Line 71; INTRODUCTION

"Although CT-guided biopsy can be effective in the case of peripheral lung tumors, tissue biopsy can be challenging in patients with early-stage operable lung cancers in clinical practice, ..."

#4 The radionics approach appears satisfactory with common use of pyradiomics and appropriate re-sampling steps.

Response: Thank you for your comments. We employed the PyRadiomics software for feature extraction which commonly uses in Radiomics community. Prior to the feature extraction, all CT images were resampled to an isotropic grid of $1 \times 1 \times 1$ mm^3 using B-splines interpolation method.

#5 It is not stated how the authors chose the feature reduction and model approaches.

Response: Thank you for your comment. Please see our response #1.

#6 The authors compared the addition of radionics with and without clinical features. The addition of radionics did not significantly improve AUC. The contribution of the clinical demographics was not otherwise quantified.

Response: Thank you for your comment. We agree with your opinion. Thus, we newly created Clinical model using all clinical variables (Age, Gender, Smoking history, and Clinical stage) and compared its predictive

performance to the Radiomics and Combined models (Table S5). As a result, the mean AUC of the Clinical model was comparable to that of Radiomics model (0.77 vs. 0.75). Moreover, although we created the All-combined model by integrating all clinical variables and radiomic features, the predictive performance did not outperform the original Combined model which includes Smoking history, Gender, and radiomic features (AUC; 0.82 vs. 0.83 as shown in Table S5). It implies that the selection of the two clinical variables is reasonable, although the Combined model did not statistically outperform the Radiomics model.

We have added some sentences and Table S5 to the revised manuscript as follows:

Page 12, Line 209; DISCUSSION

"To quantify the contribution of these clinical variables more clearly, we further created the Clinical model and the All-combined model which includes all clinical variables (age, gender, smoking history, and clinical stage) and radiomic features (Table S5). Consequently, the mean AUC of the Clinical model was identical to that of the Radiomics model (0.77 vs. 0.75). However, the predictive performance in the All-combined model was limited (i.e., the mean AUC of 0.82), implying the selection of gender and smoking history for the Combined model is reasonable."

#7 The training and validation sets are single/same centre respectively and small.

Response: Thank you for your comment. We agree with your opinion, thus we have noted it as limitation (Page 13, Line 252). Our models should be further validated with a large sample, multi-national, and multi-ethnicity datasets, and those must be future work.

#8 Table 1 - EGFR wild should state wild type

Response: Thank you for pointing it out. This has been corrected. Moreover, the term of "c-stage" has been changed to "Clinical stage".

#9 NB This study may be better suited to the TRIPOD guidelines as it has developed a model (classifier) to delineate EGFR positivity rather than demonstrating diagnostic accuracy.

Response: Thank you for your suggestion. According to TRIPOD statement, the current study is based on Type 1b. This has been added to the revised manuscript as follows:

Page 8, Line 139; METHODS

"The current study is based on TORIPOD Type 1b (22)."

Reply to Reviewer B's Comments:

Major points

#1 Because number of patients is 99, I recommend to use nested cross validation.

Response: Thank you for your comment and helpful suggestion. Unfortunately, we considered that the nested cross-validation is difficult to adopt because it needs additional test sets for outer loop in addition to the training and validation sets. Because of the limited number of patients in this study, we adopted non-nested cross-validation with Optuna. This has been added as limitation:

Page 9, Line 143; METHODS

"The performance was validated with non-nested five-fold cross-validation, ..."

Page 14 Line 269; DISCUSSION

"Fourth, we chose non-nested cross validation for evaluating the prediction performance, thus optimal hyperparameters for Random Forest were not tuned using independent validation sets."

#2"as well as clinical features (gender and smoking history)." Please describe the reason for the selection of these two variables. For example, please describe the reason for the exclusion of age and tumor markers.

Response: Thank you for your comments. We considered the two variables for Combined model because gender and smoking history are well known to clinically validated predictor of EGFR mutation [21]. Although we incorporated all clinical variables into the Radiomics model, the predictive performance was limited (the mean AUC for this model vs. Combined model were 0.82 vs. 0.83 [Table S5]).

We have added a new reference to METHODS section as follows:

Page 8, Line 137; METHODS

"Gender and smoking history were selected as these are clinically validated predictor of EGFR mutation. (21)"

#3 Details of CT examination are not described

Response: Thanks. Details of CT examination have been added to the revised manuscript and those are provided as Table S1.

Page 6, Line 98; METHODS

"All patients underwent contrast-enhanced chest CT using Discovery CT750 HD (GE Healthcare, Chicago, IL, USA). The acquisition parameters were as follows: tube voltage, 120 kVp; tube current, 100−649 mA; exposure time, 400−699 ms; milliampere-seconds, 50−329.9 mAs, slice thickness, 1.25 mm (Table S1). In ~~preoperative enhanced chest CT~~ axial CT images under lung field view ~~with 1.25 mm thickness~~, ..."

#4 It seems that number of cut-off points is too small in Figure 2. Why?

Response: Thank you for pointing this out. This isbecause the number of patients in each validation set was limited (approximately 20 patients). However, the ratio of EGFR mutation and EGFR Wild type was well balanced (almost all, 1 : 1).

#5 For reproducibility, I recommend to disclose authors' source code.

Response: Thank you for your comment. If disclosing source cord is not mandatory for publication, we would like keep it confidential as this code potentially will be used commercially.

Minor points

#1 "If the created VOI went outside the lungs (i.e., the VOI partially included mediastinum or chest wall), the contour was manually modified to fit within the lung field." If so, please clarify that there was no lung cancer that invaded mediastinum or chest wall.

Response: Thank you for your comments. We have added the description as follows:

Page 9, Line 154; RESULTS

"There was no lung cancer that invaded mediastinum or chest wall.

#2 "Using PyRadiomics version 3.0x.(13)" Please clarify the detailed version number.

Response: We have added the detailed version number for PyRadiomics.

Page 7, Line 117; METHODS

"... using PyRadiomics version 3.0.1.(13)"

#3 "Boruta algorithm for feature selection" Please cite these papers for Boruta.

https://www.jstatsoft.org/article/view/v036i11

https://pubmed.ncbi.nlm.nih.gov/34687853/

https://link.springer.com/article/10.1186/1471-2105-15-8

Response: Thank you. We have added these references to the revised manuscript [17-19].

#4 Please describe the software names of random forest and Boruta.

Response: Thanks. We have added the software names of random forest and Boruta:

Page 8, Line 129-131; METHODS

"The machine learning-based radiomics model (using the Boruta algorithm [17-19] [BorutaPy version 0.3] for feature selection and Random Forest [Scikit-learn version 1.0.2] for prediction) was constructed to detect EGFR mutation."

#5 Hyperparameters of random forest and parameters of Boruta should be described.

Response: Thank you for your comment. Some descriptions have been added to the revised manuscript as follows:

Page 8, Line 132; METHODS

"Suitable hyperparameters for Random Forest were determined by Optuna version 2.10.0 (20), while those for Boruta were empirically determined. The tuned hyperparameters by Optuna were max_depth, max_leaf_nodes, min_samples_leaf, and n_estimators. The meanings of these parameters are described in Kursa MB et al (17). The number of maximum iterations to select the optimal features for Boruta was set to 500."