Peer Review File

Article information: http://dx.doi.org/10.21037/atm-20-6473

Reviewer A

I would like to congratulate the authors for their great effort conducting the present study. A CT-derived radiomics approach for predicting primary commutations involving TP53 and EGFR in patients with advanced lung adenocarcinomas (LUAD) is described. The results are promising and may deliver a valuable tool for detecting such mutations non-invasively avoiding additional costs of expensive genetical analysis. The limitations of the study were well addressed by the authors. However, the manuscript needs thorough language and grammatical revisions. Below you'll find some suggestions for improving the manuscript.

Title Concerned -> involving

49 concerned -> involving

53 obtained -> collected

69 concerned -> involving

106 Till now -> so far

107 concerned -> involving

121 explore the alternative tool -> identify an alternative tool

125 medical image -> medical imaging

131 in this study, we aim -> delete „in this study“, we aimed

133 and 134 pretreatment images, which had not been reported yet -> pretreatment

images. To the best of our knowledge, this approach has not been previously reported

138 were obtained -> were collected

145 patients who ever received treatment -> patients previously treated

146 Patient enrollment process was detailed in -> patient enrollment algorithm is shown in detail in Figure 1.

147 Cohort clinical characteristics for analysis was given in -> cohort clinical characteristics are demonstrated in Table 1.

166 for evaluation (Table 2) were included -> for evaluation included (Table 2)

189 as given in Figure 1 -> as shown in Figure 1

191 All enhanced CT image -> images

192 to get whole -> to obtain whole

237 The features with p value exceeded 0.05 will be -> the features with p value above 0.05 were excluded

242 all the three steps -> all three steps

261 were more than 60 years -> were older than 60 years

272 contained three ones -> contained three features (or classifiers)

275 their performance were compared -> was compared

285 achieved when differentiate -> achieved in differentiating

307 concerned TP53 -> involving

318 mutations concerned TP53 -> mutations involving

321 merely compared -> merely comparing

322 for the reported, they focused -> Previous reports focused

329 mutations concerned -> mutations involving

330 therefore, extend the potential -> therefore, our model extends the potential

332 by now there is -> so far there is

364 status concerned -> status involving

Response: We thank the reviewer for the critical suggestions and valuable comments. We revised the above sentences in our article according to the reviewer's suggestion.

Reviewer B

This is an interesting study. The paper is generally well written and structured. However, in my opinion the paper needs some revisions.

Discussion: Please add an explanation for the importance of detecting EGFR and p53 mutation in a clinical settings.

Response: We strongly agree with the reviewer's opinions and make corresponding modifications to the article.

Changes in the text: Line 320-323

Reviewer C

The three subtypes of EGFR and TP53 were classified by the machine learning algorithms. The subtype classification was interesting. However, several flaws were found in this paper.

Major point

Q1: Figure 2 shows that 185 patients were used for feature selection. According to the main text, 199 were included in this study. What is this discrepancy?

Response: At the beginning of our study, 185 patients were enrolled, and then 14 patients with sequencing results were added. Finally, a total of 199 patients were included. Figure 1 is the primary flow chart and is the uncorrected number after adding cases.

Changes in the text: Revised Figure 1.

Q2: Figure 2 shows that feature selection was performed before 80% dataset were selected for training data. Feature selection should be performed after 80% dataset were selected, and the selected feature was fixed in both training dataset and test dataset. The authors' data processing led to overfitting.

Response: Thanks for the review's suggestion. Actually, we performed feature selections after the 80% dataset selection as you said, but the corresponding figure was wrongly shown. We described this information in the previous manuscript (in section 'Materials and Methods', ' Prediction Models and Workflow', paragraph 1, page 10-11). And all feature selections, classifiers establishment were based by the data in the training dataset to ensure independence from testing dataset. We modified Fig. 2 (see Figure2_revised) to fix this error.

Changes in the text: Revised Figure 2.

Q3: The numbers of selected features were 5 for all models. Why? It seems that the numbers of selected features were not specified in the procedure of feature selection.

Response: As our research samples are limited, we selected the first five features to avoid overfitting. We described this information in the previous manuscript (in section "Materials and Methods", " Feature selection", paragraph 1, page 11-12), and the nonzero feature coefficients ranking the first five were selected for each binary classifier to avoid overfitting.

Q4: In my opinion, the AUCs of radiomics model for classification between (i) EGFR+&TP53+ and EGFR- and (ii) EGFR+&TP53- and EGFR- was better than those of radiomics model for classification between EGFR+ and EGFR- reported in previous studies. As shown in my comment, I speculate that the overfitting caused by feature selection occurred in this study.

Response: Thanks for the review's suggestion. We performed feature selections in the training dataset to ensure independence from the testing dataset as shown in Response to Q3.As the reviewer mentioned, there are plenty of studies focusing on the classification between EGFR+ and EGFR- using radiomics model, but these models demonstrated various performance. Some studies showed lower AUCs than ours, e.g. Wang et al.[1] investigated the TMB status and driver mutations by radiomics features, which yielded a median AUC of 0.606, 0.604 and 0.586 respectively. While another study conducted by Jia et al.[2] identified the EGFR mutations status using random forest model with 94 radiomics features reached an AUC of 0.802, which is comparable to ours. Radiomics analysis is a promising tool that demonstrates a wide range of clinical application. However, radiomics features are affected by CT scanner types, reconstruction kernels, feature extraction software, resulting in difficulty to reproduce and validate. Feature standardization process is needed as presented by Zwanenburg's work [3] to extend its clinical use.

1. Decoding tumor mutation burden and driver mutations in early stage lung adenocarcinoma using CT-based radiomics signature. Thorac Cancer 2019 10;1010(10).

2. Identifying EGFR mutations in lung adenocarcinoma by noninvasive imaging using radiomics features and random forest modeling. Eur Radiol 2019 Sep;299(9)

3. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 2020 05;2952(2).

Q5: "Patients were categorized into three subtypes according to the mutation status of EGFR and TP53:" I cannot accept the categorization of three subtypes. Why are two subtypes ((i) EGFR- & TP53- and (ii) EGFR- & TP53-) not used? There is no explanation for use of the three subtypes instead of the four subtypes.

Changes in the text: Line 152-155.

Q6: In addition to clinical model, semantic model, radiomics model, and integrated model, I recommend to build and evaluate the following three models: (i) clinical and

semantic model, (ii) clinical and radiomics model, and (iii) semantic and radiomics model.

Response: We thank the reviewer for the critical suggestions and we think that the model classification suggested by the reviewer is reasonable. However, the model classification in our study was based on some similar studies such as [1, 2] in which the models were compared among clinical model, semantic model, radiomics model, and etc. So we believe that the model classification in our study is also reasonable.

1.Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning. Eur Respir J 2019 03;533(3).

2. A Combined Nomogram Model to Preoperatively Predict Histologic Grade in Pancreatic Neuroendocrine Tumors. Clin Cancer Res 2019 01 15;252(2).

Q7: Regarding to prediction of EGFR and TP53, please compare authors' study with the following paper: https://onlinelibrary.wiley.com/doi/full/10.1111/1759-7714.13163 .

Changes in the text: Line 344-359.

Q8: "Statistical metrics, including accuracy, sensitivity, specificity, precision, recall, F1 score, were also calculated to evaluate the overall performance of the multiclass classifier." Because of multiclass classification, please define these metrics clearly.

Response: Thanks for the reviewer's suggestion. The basic definition of these statistical metrics for multiclass classification is similar to that for binary classification, i.e., accuracy = (TP+TN)/(TP+FN+FP+TN), sensitivity = TP/(TP+FN), specificity = TN/(TN+FP), precision = TP/(TP+FP), recall = TP/(TP+FN), F1 score = 2·(precision·recall)/(precision+recall), where TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative. For multiclass classification, we supposed there are three classes: C1, C2, C3. Once we picked up one category as positive, the other two are automatically negative. For instance, 'TP of C1' means gold standard C1 instances are classified as C1. 'TN of C1' means all non-C1 instances that are not classified as C1 . 'FP of C1' means all non-C1 instances that are classified as C1. 'FN of C1' means gold standard C1 instances are not classified as C1. We added detailed description of metrics definition (in section 'Materials and Methods', ' Statistical Analysis', paragraph 1, page 12) in the revised manuscript.

Changes in the text: Line 258-260

Minor points

“Scan Protocol” is located before “Semantic CT Characteristics”.

Changes in the text: Line 163-193.

“Matpotlib” Typo. Matplotlib.

Changes in the text: Line 254.

Hyperparameters of machine learning algorithms are not shown. Please clarify them.

Response: Thanks for the reviewer’s suggestion. We compared several machine learning algorithms and found that the SVM algorithm performed best. Then, SVM algorithm was applied to the analysis and validation of multiclass classification and we list the main hyperparameters of SVM. In all the SVM algorithms, the Radial basis function kernel, also called the RBF kernel, were utilized for three base binary classifiers and four models. For each SVM, gamma and C values are the key hyperparameters that are optimized by 5-fold-cross-validation, where gamma controls the distance of the influence of a single training point and C adds a penalty for each misclassified data point. We added detailed descriptions in section ‘Results’ (in section ‘Feature Selection and The Performance of Base Binary Classifiers’, paragraph 2, page 13-14) and Supplementary Table 1.

Changes in the text: Line 286-288; Supplementary Table 1.

Reviewer D

I reviewed the paper entitled “A CT-derived Radiomics Approach for Predicting Primary Commutations Concerned TP53 and EGFR in Patients with Advanced Lung Adenocarcinomas (LUAD)”. This paper tries to effectively predict the three subgroups of gene mutation status in advanced NSCLC patients using radiomic features extracted from CT imaging.

To my opinion the paper has some power points but simultaneously it has some weakness that should be addressed before it can be considered for publication. Below are my major comments:

Q1: Why did you decide to divide your cohort to 80% for training and only 20% for validation? In this situation, you will have only 40 patients in validation set. Although, it is not clear how many patients do you have for each 3 subgroups, the precision of

your model cold be controversial.

Response: Thanks for the reviewer's suggestion. As the total number of our study cohort is limited, we have to obtain sufficient samples for training to guarantee the robustness of the model construction, so we divided our cohort to 80% for training and 20% for validation. We have mentioned the limitation in section 'Discussion' (paragraph 5, page 18, 'it still has several limitations. First, our findings deserve further study with expanded samples and extra external validation.')

Changes in the text: Line 264-269.

Q2: Please indicate the number of mutations in training and validation sets.

Changes in the text: Line 264-269.

Q3: Please confirm that if the results in the result sections are for training set or validation set?

Response: The results in Table 4 are for the training set, and results in Table 5 and figure 3 are for the validation set. We revised the two sentences in the manuscript as 'We selected the SVM algorithm with the best performance for the training dataset as shown in Table 4, and all the analysis and validation were based on it. ' (in section 'Results', 'Feature Selection and The Performance of Base Binary Classifiers', paragraph 2, page 13) and 'The performance of the multiclass classification for the validation dataset, in our study was three-type classification, was shown in Table 5 and Figure 3. ' (in section 'Results', 'Multiclass Classification Strategy', paragraph 1, page 14).

Changes in the text: Line 285-286, 301.

Q4: Two decimals for AUCs are enough.

Response: Thanks for the review's suggestion. Due to three decimals for P value in Table 1 and Table 2, and in order to maintain the data consistency, we also used three decimals for AUCs, sensitivity as well as specificity. Considering the potential errors that may occur during the modification process, this article still uses three decimals for AUCs.

Q5: The discussion section needs to be revised significantly. Don't repeat your result section again. You should discuss what radiomic features were able to predict mutation status and more importantly, what is the rationale behind it. In other words, why

radiomic features extracted from CT can predict the mutation status.

Response: Thanks for the reviewer's suggestion. Since most of the radiomics features represent the pixel arrangement and gray level of the image, most of them cannot be explained clinically. Therefore, the rationale behind them is very difficult to explain. The previous literature on artificial intelligence did not explained these features also. We modified the discussion part, emphasizing the clinical significance of this study, and compared with the existing studies.

Changes in the text: Line 316-343.

Q6: I think it would be better to divide patients in table 1 and 2 into training and validation sets and also don't forget to report the p-values.

Response: Thanks the reviewer for the critical suggestions.

Changes in the text: Line 264-278, Revised Table 1, Revised Table 2.

Q7: I think there are a lot of other works that have evaluated the gene expression on CT images. Please cite them and discuss in the discussion section.

Changes in the text: Line 330-359

Q8: I'm not sure about the limited number of references in this journal but I think 24 references should be very low.

Changes in the text: Line 415-554.

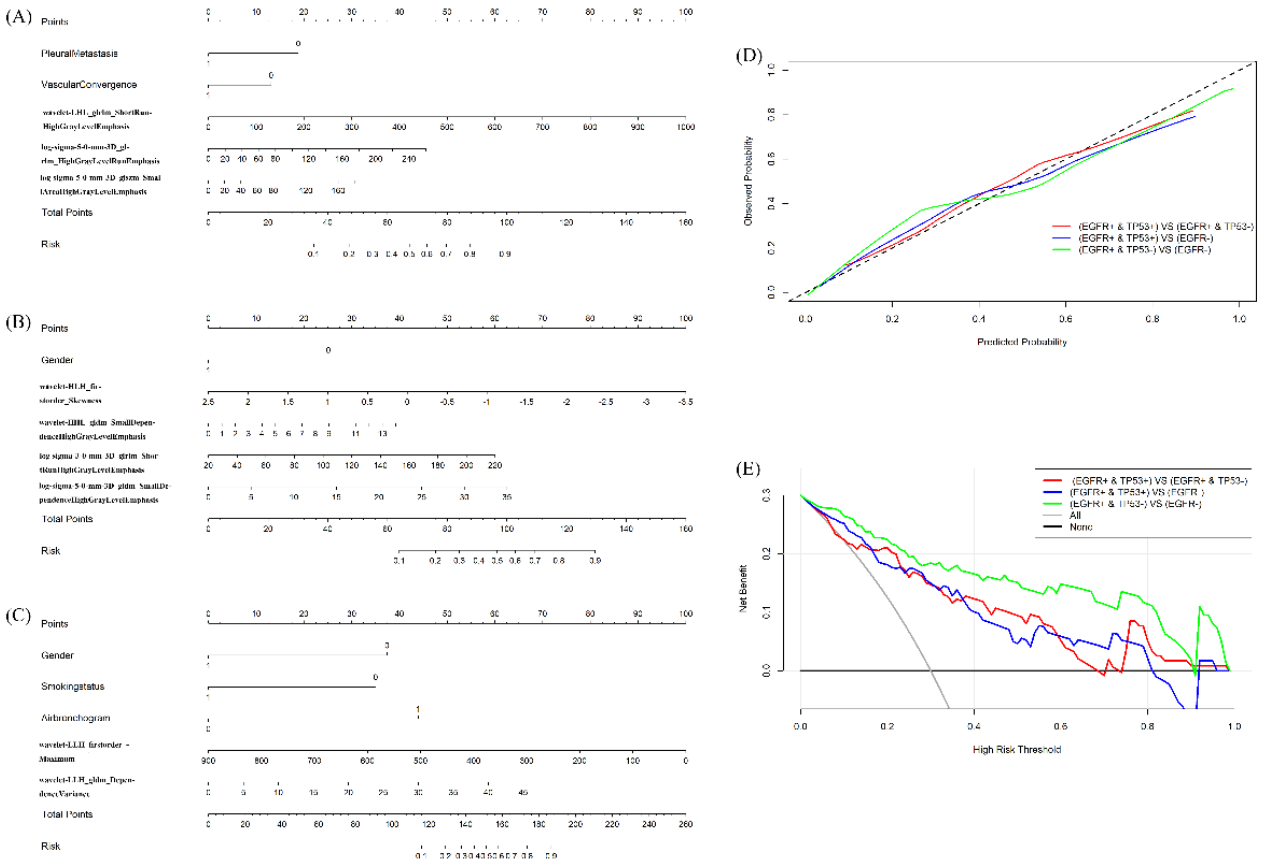
Reviewer E

This study addressed the AI-based radiomic features could help in differentiation of the status of TP53(positive or negative) and EGFR (positive and negative). Therefore, this way could help the target therapy without tissue sampling. However, no validation cohort design in this study. This part is quite important, in the development of predictive models (discrimination, validation, calibration).

1.Please develop the individual monogram for clinical practice in predicting the status of TP53 (positive or negative) and EGFR (positive and negative)

Response: The reviewer's suggestion is greatly appreciated. The nomogram provides a practical prediction means for clinical usage. However, this method is mainly developed for binary classification, and may be not suitable for the multiclass

classification in our study. Since we developed three base binary classifiers as the intermediate step, the nomograms of the integrated model for these three base classifiers were shown below:



Development and performance of the nomograms of the integrated model for three base binary classifiers ((EGFR+ & TP53+) VS (EGFR+ & TP53-); (EGFR+ & TP53+) VS (EGFR-); (EGFR+ & TP53-) VS (EGFR-)). (A) Nomogram for the classification of (EGFR+ & TP53+) and (EGFR+ & TP53-). (B) Nomogram for the classification of (EGFR+ & TP53+) and (EGFR-). (C) Nomogram for the classification of (EGFR+ & TP53-) and (EGFR-). (D) Calibration curves for three base binary classifiers. (E) Decision curve analysis for three binary classifiers.

2. Please add the study outcome diagnostic flowchart because of complex classification in this study design (TP53, positive or negative and EGFR, positive and negative).

Response: The study outcome diagnostic flowchart is shown in detail in Figure 1. Patients in the training cohort and the validation cohort were shown in Revised Table 1 and Table 2.

Changes in the text: Line 264-278, Revised Table 1, Revised Table 2.

3. In the introduction part, please address the high prevalence of non-smoking related lung cancer with associated with EGFR/TP53 status in Asian population.

Therefore the audience could understand the clinical indication of this paper.

References:

1. Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell Volume 182, Issue 1, 9 July 2020, Pages 226-244.e17

2. Wu FZ, Huang YL, Wu CC, Tang EK, Chen CS, Mar GY, et al. Assessment of Selection Criteria for Low-Dose Lung Screening CT Among Asian Ethnic Groups in Taiwan: From Mass Screening to Specific Risk-Based Screening for Non-Smoker Lung Cancer. Clinical lung cancer 2016; 17: e45-e56, doi:10.1016/j.cllc.2016.03.004.

3. Hsu HT, Tang EK, Wu MT, Wu CC, Liang CH, Chen CS, et al. Modified Lung-RADS Improves Performance of Screening LDCT in a Population with High Prevalence of Non-smoking-related Lung Cancer. Academic radiology 2018; 25: 1240-1251, doi:10.1016/j.acra.2018.01.012.

4. Zhou F, Zhou C. Lung cancer in never smokers-the East Asian experience. Transl Lung Cancer Res 2018; 7: 450-463, doi:10.21037/tlcr.2018.05.14.

Response: Thanks the reviewer for the critical suggestions.

Changes in the text: Line 90-96.

4. Grammar and spelling

A few minor typos, grammar hiccups should also be corrected, e.g., “component surround” in line 173, page 8.

Page 8 -Corrected grammatical mistake in the “component surround” to “component surrounded”” (line 173)

Page 9 -Corrected grammatical mistake in the “used as contrast agent” to “used as the contrast agent”” (line 185)

Page 9 -Corrected grammatical mistake in the “25 years experiences.” to “25-year experience.”” (line 196)

Page 15 -Corrected grammatical mistake in the “have better clinical outcome” to “have a better clinical outcome” (line 315)

Page 17 -Corrected grammatical mistake in the “Forth” to “Fourth” (line 360)

Response: We thank the reviewer for the critical suggestions and valuable comments. We revised the above sentences in our article according to the reviewer's suggestion.