Peer review file

Article information: http://dx.doi.org/10.21037/jtd-21-80

Reviewer A

Comment 1: Mean +/- SD was mentioned in 'Statistical analysis' (line 145), but you use mean +/- SE in Table 1. Please check again and unify.

Reply 1: Thanks a lot for the careful evaluation. All the numerical data were presented with mean ± standard deviation in the current study. We have corrected the note to ***Table 1***.

Changes in the text: Page 24, line 444.

Comment 2: What is the percentage of missing data for CEA? Please mention because imputation with a single number (median) may not be appropriate if missing percentage is not low. There should be much better to perform the missing data imputation.

Reply 2: Thank you very much for the reminder. The percentage of missing data for

CEA was 0.26 in malignant group and 0.41 in benign group. Details of missing data for CEA, CYFRA21-1 and NSE were also summarized in ***Table S3***.

Changes in the text: See **Page 11, line 200-201** and ***Table S3*** in supplementary material.

Comment 3: Why did you perform 1:1 matching rather than include all nodules in SSN analysis? You also mentioned that this approach reduced the sample size and made the model unstable (line 227-228). And, the prediction of proposed SSN model is not very well compared with others.

Reply 3: We totally enrolled 623 SSNs, and the ratio of 519 malignant SSNs to 104 benign SSNs was 5:1. There were much more positive samples (malignant) than negative samples (benign).

When building models, especially the machine learning model, positive and negative sample imbalance will lead to overfitting. The prediction can be biased to the classification that harbors more samples, for example, the malignant classification in the current study. Consequently, the generalization capability of models would be

impaired (1).

Here, we tried to build SSN models based on all nodules. From the results we found the AUC of the models wasn't improved and the specificity became tremendously low (over fitting and bias). The results on validation dataset were showed in the following Table 1 and Figure 1.

Thank you for your comment. We've added the reason why we matched the sample in the manuscript.

Changes in the text: Page 11, line 205-207.

Table 1 Performance of SSN models on validation dataset.

Samples	Methods	AUC	Sensitivity	Specificity
1:1 matching	XGBOOST	0.65	0.64	0.60
	Random Forest	0.73	0.73	0.67
	Logistic	0.76	0.69	0.76
All SSNs	XGBOOST	0.64	0.90	**0.29**
	Random Forest	0.71	0.79	**0.58**
	Logistic	0.77	0.98	**0.13**

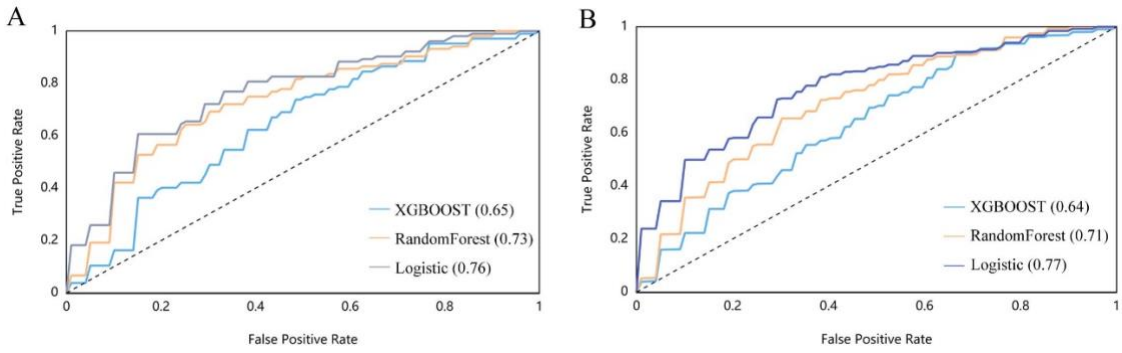

Figure 1 ROC curves of SNN models. A. 1:1 matching; B, all SSNs.

Comment 4: You mentioned clinical features account for 60% of top features, and also mentioned VDT as an important feature. Did you use it in your model?

Reply 4: This was an interesting question. We didn't use volume doubling time (VDT) to build models in the current study. Usually at least two thoracic CT scans prior to treatment are needed to calculate the VDT of a nodule, but most of patients we studied didn't meet the criteria.

Actually, we have investigated the growth characteristics of benign and malignant pulmonary nodules in a small sample size (N=305) (2). However, the results suggested that the diagnostic value of VDT was limited, with AUC being 0.67 (95%CI, 0.55–0.78), sensitivity and specificity being 69% and 58%, respectively (2).

Nevertheless, the current literatures and guidelines do imply that the growth of

pulmonary nodules is associated with lung cancer probability(3-5). Hence, at present we are doing another research to better characterize the growth characteristics of pulmonary nodules based on large sample size and optimized study design. Hopefully, we can not only establish models to predict the likelihood of malignancy but also provide scientific evidence in relation to follow-up strategies of pulmonary nodules.

Changes in the text: We clarified this point in the limitation part of Discussion (**Page 16, line 314-318**).

Comment 5: In exclusion criteria, what is the definition of multiple pulmonary nodules, and why did you decide to exclude it? Because it is very common to detect 'more than one' nodules by HRCT, this exclusion criteria will really limit the model application in the clinical practices.

Reply 5: Thank you very much for the comment.

From guidelines, pulmonary nodules are defined as focal opacities that measure up to 3 cm in diameter and are surrounded by lung parenchyma, including those abutting the pleura (6). In the individual with a dominant nodule and one or more

additional small nodules, it is referred to as multiple pulmonary nodules (7).

In the current study, the patients we excluded with multiple pulmonary nodules mainly meet the following criteria: (1) nodules distributed diffusely; (2) multiple nodules were pathologically confirmed to be tumors; (3) there were several problematic nodules.

We decided to exclude these patients because (1) diffused nodules were out of our interest; (2) sometimes it was difficult to distinguish multiple primary tumors from metastatic tumors and (3) from the multiple problematic nodules it was hard to choose the right one for further analysis.

Changes in the text: We modified the exclusion criteria in relation to multiple pulmonary nodules to make it clearer (**Page 7, line 107-109**).

Reviewer B

Comment 1: Please add a supplemental Table which details the histopathological diagnosis of 1,171 patients enrolled, by categorizing following 4 groups: Solid-malignant, solid-benign, subsolid-malignant, and subsolid-benign.

Reply 1: Thank you very much for the suggestion. We've added one supplemental table describing histopathological diagnosis of all enrolled nodules in 4 groups as recommended.

Changes in the text: See **Page 10, line 190** and ***Table S2*** in supplementary material.

Reviewer C

Comment 1: There is no mention of rate of growth on the predictive values

Reply 1: Thanks a lot for the question. Usually at least two thoracic CT scans prior to treatment are needed to calculate the growth rate (volume doubling time, VDT) of a nodule, but most of patients we studied didn't meet the criteria. Hence, we didn't use VDT to build models in the current study.

Actually, we have been investigating the growth characteristics of benign and malignant pulmonary nodules. Details can be seen in Reply 4 to Reviewer A. By studying growth rate of nodules, we hope not only establish models to predict the likelihood of malignancy but also provide scientific evidence in relation to follow-up strategies of pulmonary nodules.

Changes in the text: We clarified this point in the limitation part of Discussion (**Page 16, line 314-318**).

Reviewer D

Comment 1: Poor English, language editing is needed. It should not be published without language editing.

Reply 1: Thanks for the suggestion. We have obtained technical help with English language in revising our manuscript.

Changes in the text: Professional language editing was performed for the whole text.

Comment 2: More structure needed throughout the entire manuscript, so that the scope and different steps are clearer.

Reply 2: Thank you. Done as requested.

Changes in the text: Necessary structure was added in the Methods and Results (**Page 6-13, line 100-244**).

Comment 3: Consider focusing on the objective of the study: developing of a risk model for small solid and subsolid pulmonary nodules, based on clinical and quantitative radiomics features.

Reply 3: Thanks a lot for the advice. The revised title is “Developing of risk models for small solid and subsolid pulmonary nodules based on clinical and quantitative radiomics features”

Changes in the text: Page 1, line3-4.

Comment 4: Adjust the goal of the study in the abstract and manuscript.

Reply 4: Thank you. Done as suggested.

Changes in the text: Page 3, line31-34 and **Page 6, line 95-97**.

Comment 5: Your study design has nothing to do with preventing the disease. Focus on assessing likelihood of malignancy and nodule management.

Reply 5: We really appreciate the comment. Inappropriate statement has been revised.

Changes in the text: Page 5, line 67.

Comment 6: The problem with morphological features is not only the inter- and intraobserver agreement, but the fact that morphological features overlap between benign and malignant. Morphological features are on themselves not sufficient to assess the likelihood of malignancy.

Reply 6: Thank you very much for your reminder. We've added this important point in Introduction section.

Changes in the text: Page 5-6, line 83-85.

Comment 7: 'some missing data of laboratory test were populated by a median value': please briefly explain

Reply 7: Thanks. There were some missing data regarding to CEA, CYFRA21-1 and NSE test. The data weren't available because patients didn't undergo the tests. We've added brief explanation in the manuscript.

Changes in the text: Page 7, line 126-128.

Comment 8: CT image acquisition: is this really true? That the contrast enhancement doesn't impact radiomics analysis?

Reply 8: Thank you for the question. We did a literature review and found three articles that investigated effects of contrast-enhancement on the performance of radiomics signature in lung cancer. Different results were found among studies.

Wu et al. found the AUC of the radiomics model was 0.86 both for images with and without contrast enhancement, suggesting that contrast enhancement did not impact the utility of radiomics analysis (8). However, He et al. revealed that non-contrast CT based radiomics signature demonstrated better discrimination capability than contrast-enhanced CT (AUC, primary cohort 0.862 vs. 0.829, $p = 0.032$; validation cohort 0.750 vs. 0.735, $p = 0.014$) (9). Besides, Kakino et al. found that contrast enhancement in the delayed phase of CT images for non-small cell lung cancer patients affected some of the radiomic features and the variability of radiomic features due to contrast uptake may depend largely on the patient characteristics (10).

Hence, it's relatively controversial regarding to the effect of contrast-enhancement on the radiomics signature in lung cancer. We revised the manuscript to

be more rigorous. Thanks again.

Changes in the text: Page 8, line 136-137.

Comment 9: Two specialists who were blinded to the pathological results of lesions evaluated all CT scans what did they evaluate? This is not clear. Furthermore, in the next paragraph it is stated that the nodules were segmented by one author. This was not a radiologist?

Reply 9: The two specialists evaluated CT scans to help select patients based on inclusion and exclusion criteria. Besides, they recorded the radiologic features of each nodule, such as location, shape, spiculation, lobulation and texture. We have added this information in text to make it clear.

The author who performed nodule segmentation was a clinician in Pulmonary and Critical Care Medicine, with four years of experience in pulmonary nodule evaluation. We've rewrote the sentence.

Changes in the text: Page 8, line 138-139 and **Page 8, line 144-146**.

Comment 10: Were this incidental pulmonary nodule? Screen detected nodules or a combination?

Reply 10: The current study enrolled both the incidental pulmonary nodules and annually screen detected nodules. We have modified our text to make it clear.

Changes in the text: Page 6, line 105-106.

Comment 11: Development of the risk models: the rationale and statistics behind it should be discussed in materials and methods, but the results should be discussed in the 'results' section of the manuscript.

Reply 11: Thanks for the suggestion. We've moved the results in that part to "results' section of the manuscript.

Changes in the text: Page 11, line 204-207.

Comment 12: 97.4% adenocarcinomas, 2.0% squamous and 0.7% other types: 100.1%?

Reply 12: Thank you for your careful evaluation. We checked the data and found it's better to keep two decimal places. When corrected, it's 97.37% adenocarcinomas,

1.97% squamous carcinomas and 0.66% other types.

Changes in the text: Page 10, line 187-190.

Comment 13: 'was observed between lung cancer and control groups': this is not mentioned earlier in the manuscript. Please explain in materials and methods.

Reply 13: The "lung cancer and control groups" actually refers to the "malignant and benign groups". We're sorry for the inappropriate description and have replaced it with "malignant and benign" in the text.

Changes in the text: Page 11, line 200.

Comment 14: What about the missing data?

Reply 14: Details of missing data were summarized in ***Table S3***.

Changes in the text: See **Page 11, line 200-201** and ***Table S3*** in supplementary material.

Comment 15: Don't mix data, studies and evidence of screen-detected nodules with

incidental nodules!

Reply 15: Thank you for the kind reminder. We've checked the whole Discussion and modified relevant discussion where the data, studies and evidence of screen-detected nodules mixed with those of incidental nodules.

Changes in the text: Page 14, line 268, Page 15, line 288-291 and **Page 15, line 301-303**.

Comment 16: 'can be accepted to some degree': limitations are limitations. If you believe the impact should be minimized, that should be discussed. The difference in scanner machines and protocols has a relevant impact and should not be minimized. 'Can be accepted to some degree' rather feels like an impression of the authors than a decision based on evidence and data.

Reply 16: Thank you for your comment. We should have described this shortcoming more objectively. The relevant content was revised.

Changes in the text: Page 16, line 310-314.

Comment 17: Strongly emphasize the shortcomings of the study and what will be needed in the future to solve this problem.

Reply 17: Thanks for the comment. Except for the shortcomings we've mentioned in text, there were two additional aspects were added.

Growth rate of pulmonary nodules is one of the key characteristics associated with lung cancer probability, but we failed to apply the relevant parameters to build risk models in the current study. Usually at least two thoracic CT scans prior to treatment are needed to calculate the growth rate of a nodule, but most of patients we studied didn't meet the criteria. Besides, only a small number of SSNs were available for modeling and therefore the performance of established SSN models was unstable. Increasing the sample size, especially for the benign SSNs, is warranted in future research.

We've revised the limitation part in Discussion. Thank you again.

Changes in the text: Page 16, line 314-321.