Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-23-633

Reviewer A

Summary: The manuscript entitled "Machine learning model for circulating tumor DNA detection in COPD Patients with lung cancer" presents results from the study on 177 patients with COPD and lung cancer.

In the study Authors have performed deep sequencing to genotype the ctDNA from plasma using targeted deep sequencing. The study focuses on the detection of cancer-related ctDNA mutations, clinically important variants. One of the aims of the study was the creation of a model predicting the ctDNA presence.

Comment 1. It would be helpful to provide additional explanations for readers who are not well-versed in machine learning about the specific prediction task that the model was designed to perform.

Reply 1. The machine learning model was designed to predict the "ctDNA mutation detection" as the binary outcome (detected vs. not detected) in COPD patients with newly diagnosed lung cancer. To provide detailed explanations of machine learning, we moved the paragraph from supplementary material to the main manuscript.

Changes in the text (Page 11, Paragraph 1)

"To predict whether each COPD patient harbors ctDNA mutation or not, we considered four binary classifying ML models (logistic regression (LR), elastic net logistic regression (EN), random forest (RF), and support vector machine (SV)). After splitting the dataset into training and test sets within the frame of leave-one-out cross-validation, we selected variables as features for ML models that showed significant association ($P < 0.1$) with the presence of ctDNA mutation in a univariable logistic regression model within each training set. The hyperparameters for EN, RF, and SV models were optimized by using grid search 5 cross-validation for accuracy in each training set. EN model was tuned by alpha from 0.0001 to 100, and L1 ratios between 0.0 and 1. RF model was allowed to have 10 to 1000 estimators, maximum depth between 6 and 12, minimum samples per leaf between 8 and 18, and minimum samples per split between 8 and 20. SV model was allowed to use either radial or linear kernels, with gamma and C parameters between 0.001 to 100. To evaluate each model, we estimated the area under the receiver operating characteristics (ROC) curve (AUC), accuracy, sensitivity, specificity, and positive predictive value in the test set, and represented the performance of each model using an ROC curve plot. The model with the highest AUC was selected as the best prediction model for the shedder."

Comment 2. Additionally, could you elaborate on how the model can assist in early cancer diagnosis?

Reply 2. One of the major limitations of using plasma ctDNA for early cancer diagnosis is the low sensitivity (tumors are not shedding enough amount of ctDNA into circulation), especially in early-stage cancer. Our model predicts the "ctDNA detection" in an individual patient with

COPD and lung cancer, based on their demographics, clinical characteristics, tumor characteristics, and radiologic parameters. For example, in patient whose probability of ctDNA detection is calculated as "high", then ctDNA analysis can be utilized for the diagnosis of lung cancer. On the other hand, for some patients with low probability of ctDNA detection despite that they have lung cancer, clinicians would not rely on the ctDNA in diagnostic process. Therefore, this model can assist in non-invasive lung cancer diagnosis, by providing estimation of the probability of ctDNA detection in COPD patient with suspected lung cancer. We added this to the Discussion section in the revised manuscript.

Changes in the text (Page 18, Paragraph 2)

"The strength of this study is the use of data from targeted deep sequencing and the adoption of ML models to predict ctDNA detection in an individual patient with COPD and lung cancer. The model can assist in non-invasive lung cancer diagnosis by estimating a probability of ctDNA detection. For example, based on the prediction scores, more than 82% of patients with the top 10% score had ctDNA mutations, suggesting that the diagnosis of lung cancer can be established using ctDNA in these patients. On the other hand, for some patients with low prediction scores (low probability of ctDNA detection), clinicians should also utilize other diagnostic tests rather than solely rely on the ctDNA analysis."

Comment 3. It may also be beneficial to provide more information on the relationship between ctDNA and poor OS.

Reply 3. In response to the reviewer's comment, we added more information (with more references) on the relationship between ctDNA detection and poor OS in the Discussion section in the revised manuscript.

Changes in the text (Page 17, Paragraph 3)

Regarding the prognostic value of ctDNA detection, this study confirmed the findings of previous studies by showing that ctDNA detection and VAF levels are associated with shorter OS, particularly in early stages (17,38-40). Poor survival with positive ctDNA in early-stage lung cancer might stem from the higher recurrence rate after the surgery. Numerous studies have shown that preoperative and postoperative ctDNA detection was associated with shorter recurrence-free survival (RFS) and OS after curative surgery (41-43). Accordingly, previous studies suggested the presence of ctDNA mutations from liquid biopsy into cancer staging as TNM "B" tumor staging, as ctDNA detection may reflect the presence of micrometastasis or minimal residual disease beyond a mere reproduction of information from tissues (44). Moreover, recent studies have shown that specific mutational profile or tumor mutational burden in ctDNA can also predict poor clinical outcome with polyclonal metastasis pattern and treatment response to immune checkpoint inhibitors (45,46).

Comment 4. Also, in the results section, it would be good to elaborate more on the overall detection of the ctDNA, and how the authors distinguished the presence of the ctDNA from cfDNA. What were the biological characteristics of the detected ctDNA compared to

the cfDNA (e.g.. fragment sizes, genomic position of a fragment, and fragment end motifs)?
Reply 4. In response to the reviewer's suggestion, we now elaborate more on the overall detection of the ctDNA in the Results section as follows.
Regarding how we distinguished ctDNA from cfDNA, the supplementary method provided a detailed description of ctDNA detection methods ("Detection of somatic mutations", page 19-20 of original submission file). Briefly, only reliable variants detected by the target panel sequencing were selected, and probabilistically detectable errors were defined and excluded. Of note, this study did not include the use of epigenetic information (fragment size or pattern) from cfDNA to increase the reliability of detected variants as it was considered to have only a marginal benefit. We now moved the supplementary methods to the main manuscript.

Changes in the text (Page 12, Paragraph 2)
*"Among the 177 patients with COPD and treatment-naïve lung cancer, at least one ctDNA mutation was detected in 54 (30.5%) patients. Detection rate was 8.1%, 25.8%, 52.8%, and 63.2% in stage I, II, III, and IV, respectively. The median (range) number of detected mutations per patient was 2 (1 – 8) and the median (range) VAF of the mutations was 6.0% (0.7% – 85.3%). The most frequently mutated genes were TP53 (70%), RB1 (19%), CSMD3 (15%), KEAP1 (9%), and LRP1B (9%) (**Figure 1A**). TP53 was the most frequently mutated gene in both adenocarcinoma (52.9%) and squamous cell carcinoma (69.6%)."*

Changes in the text (Page 9, Paragraph 2)
"First, all bases were subjected to Phred quality filtering using a threshold Q of 30 and only positions where total depths were above 500× were considered for variant identification. To exclude germline mutations in the analysis, non-reference alleles present at a frequency greater than 1% in the matched white blood cell gDNA were removed. The error suppression method using UMIs was used to distinguish true somatic mutations from PCR and sequencing errors. After applying the error suppression method to the sequencing data, the following selection steps were used to eliminate the remaining sequencing errors: (1) variants not significantly greater than the error found in the matched germline DNA (binomial Bonferroni-adjusted $P < 0.01$) were filtered out; (2) variant candidates with a high strand bias (90% if supporting reads ≥ 20; Fisher's exact test, $P < 0.1$ if supporting reads < 20) were removed; (3) if the z-statistic of the variants was not significantly higher than the background error obtained from gDNA (Bonferroni-adjusted $P < 0.05$), they were excluded from the analysis.
Finally, the mutation candidates were selected according to the following conditions: Allele frequencies $\geq 0.15\%$ and alternative allele counts ≥ 5 were selected. For tissue specimens, somatic variants were identified using different criteria: total depth $\geq 100\times$ and allele frequency $\geq 2\%$. In the case of insertions or deletions, variants with an allele frequency $\geq 5\%$ were selected. Variants were annotated using VEP (v102) (23) and nonsynonymous variants were used in this analysis."

Comment 5. Regarding the section titled "Clinical factors associated with ctDNA detection", it is mentioned that ctDNA was detected among 54 patients. Therefore, when the subsequent paragraphs refer to patients with ctDNA presence, are they referring to

the same group of 54 individuals?

Reply 5. Yes, you are correct. To clarify, we added "(N = 54)" after "patients with ctDNA mutations".

Changes in the text (Page 13, Paragraph2)

"As a result, patients with ctDNA mutations (N = 54) had higher mMRC grade, lower EI, higher CRP, larger tumor size, more advanced clinical stages, more centrally located tumors, and a higher prevalence of small cell lung cancer (SCLC) than patients without ctDNA detection (Table 1)."

Comment 6. Could you provide more details about the emphysema index? Is it feasible to merge the analysis of the emphysema index with ctDNA genotyping to detect lung cancer early in patients with COPD or those at high risk of developing it?

Reply 6. Emphysema index is the percentage (quantitation) of lung area with CT attenuation values < −950 HU (low attenuation area) in the whole lung at inspiration, which is automatically measured using software program. It is known that the presence and severity of emphysema is associated with the risk of lung cancer (Radiology 2022; 304:322-330). Therefore, as the reviewer suggested, emphysema index can be merged with ctDNA genotyping and other individual risk factors such as smoking history to detect lung cancer early in patients with COPD. However, in this study of patients with COPD and lung cancer, we found that the higher emphysema index is also associated with the lower probability of ctDNA detection, which might complicate the utility of ctDNA genotyping in patients with severe emphysema. Therefore, the utilizing both variables into lung cancer prediction model needs further separate study for comparison between COPD patients with lung cancer and those "without lung cancer", which will be the next step for the use of ctDNA genotyping.

Changes in the text: None.

Comment 7. In the Reviewer's opinion, the study population description should be included in the main manuscript. Table 1, with the characteristics of the COPD population, should also be cited in the method section.

Reply 7. In response with the reviewer's suggestion, we now moved the description of the study population from supplementary material to the main manuscript. Table 1 and the characteristics of the study patients were in the Results section and we would like to keep them in the Result section. However, we moved the paragraph to the beginning of the Results section in the revised manuscript.

Changes in the text (Page 8, Paragraph 4)

"From October 2017 to September 2020, 461 patients with spirometry-defined COPD (post-bronchodilator $FEV_1/FVC < 0.7$) aged ≥ 40 years were prospectively enrolled from a single referral hospital. All of them did not have a significant pulmonary fibrosis. After excluding patients whose blood samples did not pass quality control or had technical issues in sample

processing or library preparation (N = 43), who withdrew consent or did not collect blood samples (N =11), and who had a history of malignancy other than lung cancer (N = 3), 404 patients were included in the study population. For the present study, we further excluded patients without lung cancer (N = 209), those with missing variables (N = 10), and never smokers (N = 8). Finally, 177 COPD patients with newly diagnosed lung cancer were included in the analysis."

Changes in the text (Page 12, Paragraph 1)

*"The clinical variables of patients with COPD and treatment-naïve lung cancer (N = 177) are summarized in **Table 1**. Overall, the mean (standard deviation, SD) age was 69.8 (6.7) years, and most patients were male (94.8%) with former (67.2%) or current (32.8%) smoking exposure. Symptomatic burden measured using the mMRC dyspnea scale and CAT was relatively mild. Pulmonary function tests showed that the mean FEV_1 was 70.3% pred, and the majority of patients (90.4%) had $FEV_1 \geq 50\%$ pred. The mean (SD) EI was 4.43 (6.81) % and 24 (13.6%) patients had an EI ≥ 10%. The clinical stages of lung cancer were classified as stage I (41.8%), II (17.5%), III (29.9%), and IV (10.7%), respectively. There were eight patients whose lung cancers were not histologically confirmed, mainly due to poor lung function, although the clinical diagnosis of lung cancer was unequivocal. Most patients (94.0%) had non-small cell lung cancer (NSCLC) and adenocarcinoma accounted for 40% of NSCLC cases."*

Comment 8. While performing the correlation analysis and the multivariable analysis, did the Authors take into consideration the presence of the node involvement?

Reply 8. We did not take into consideration the presence of the node involvement in the analysis. However, we have TNM stage of lung cancer (from I to IV), which encompasses the nodal involvement.

Changes in the text: None.

Reviewer B

Comment 1. Did the patients have concomitant fibrosis?

Reply 1. All of the patients included in this study had a post-bronchodilator spirometry confirmed COPD and those with clinically significant pulmonary fibrosis were not included from the screening step. We added this in the Methods section of the revised manuscript.

Changes in the text (Page 8, Paragraph 4)

"From October 2017 to September 2020, 461 patients with spirometry-defined COPD (post-bronchodilator $FEV_1/FVC < 0.7$) aged ≥ 40 years were prospectively enrolled from a single referral hospital. All of them did not have a significant pulmonary fibrosis."

Comment 2. Were they current smokers or former smokers?

Reply 2. As shown in the Table 1, all of the patients were current (32.2%) or former (67.8%) smokers. This was also mentioned in the Results section of the original manuscript (Page 5, Line 195-196 of the original submission).

Changes in the text: None.

Comment 3. history of occupational exposure?

Reply 3. Unfortunately, our data lack occupational exposure in this prospective cohort of COPD patients with newly diagnosed lung cancer. Therefore, we do not know whether or not the presence of the occupational lung cancer is associated with the circulating tumor DNA detection. Similarly, as we only included smokers in this analysis, the same applies to the smoking exposure. We added this limitation in the Discussion section of the revised manuscript.

Changes in the text (Page 18, Paragraph 3)

"In addition, as this study focused on COPD patients recruited from pulmonology clinics, our cohort predominantly consisted of men and smokers (> 90%), which is consistent with the multicenter studies from Korea that are based on pulmonology clinics (40,41). This may limit the generalizability of our findings to other populations. As all patients were current or former smokers, the relatively lower prevalence of EGFR mutations in our cohort might be attributable to smoking and COPD (35,42). The lack of never-smokers made it impossible to explore the association between smoking exposure and ctDNA detection. Similarly, due to the unavailability of occupational information, the association between occupational exposure and ctDNA detection was not investigated."