Peer Review File

Article information: https://dx.doi.org/10.21037/tlcr-23-473

Reviewer A

Given the small study population I see no reason to have one training/validation split as opposed to bootstrapping the entire procedure. Now your validation split consists of only 10 patients. Consider LOOCV. Also, strictly speaking, ANY data transformation that is dependent on the data distribution should be performed independently in train/validation. I.e. the expression normalisation using the housekeeping genes should take place as part of the cross-validation.

Reply 1:

Thank you for pointing this out. We agree with this comment. The size of the study population is the key limitation in our study, in the preliminary ML model developing stage, we tried several validation methods including leave one out, k-fold, and repeat cross validation, among those CV methods, repeat cross validation showed the best performance on our dataset, that's why we use repeat CV in all of our tumor tissue, buffycoat and clinical datasets. And yes, the data normalization and transformation should be performed absolutely independently in training and validation steps, and that was how we did it.

Housekeeping normalisation: Reasonable approach but please indicate the housekeeping genes explicitly or refer to the relevant work.

Reply 2:

Thanks for this suggestion. Since we used nSolver to do the normalization, the housekeeping genes were given by this software (listed below). We added this information to the "methods" section as supplemental table 1.

Changes in Line172-173: "The list of the housekeeping genes was provided in Supplemental Table 1."

Normalization Codes

Probe Name	Class Name
ABCF1	Housekeeping
AGK	Housekeeping
ALAS1	Housekeeping
AMMECR1L	Housekeeping
CC2D1B	Housekeeping
CNOT10	Housekeeping
CNOT4	Housekeeping
COG7	Housekeeping
DDX50	Housekeeping
DHX16	Housekeeping
DNAJC14	Housekeeping
EDC3	Housekeeping
EIF2B4	Housekeeping
ERCC3	Housekeeping

Normalization Codes

Probe Name	Class Name
FCF1	Housekeeping
G6PD	Housekeeping
GPATCH3	Housekeeping
GUSB	Housekeeping
HDAC3	Housekeeping
HPRT1	Housekeeping
MRPS5	Housekeeping
MTMR14	Housekeeping
NOL7	Housekeeping
NUBP1	Housekeeping
POLR2A	Housekeeping
PPIA	Housekeeping
PRPF38A	Housekeeping

Normalization Codes

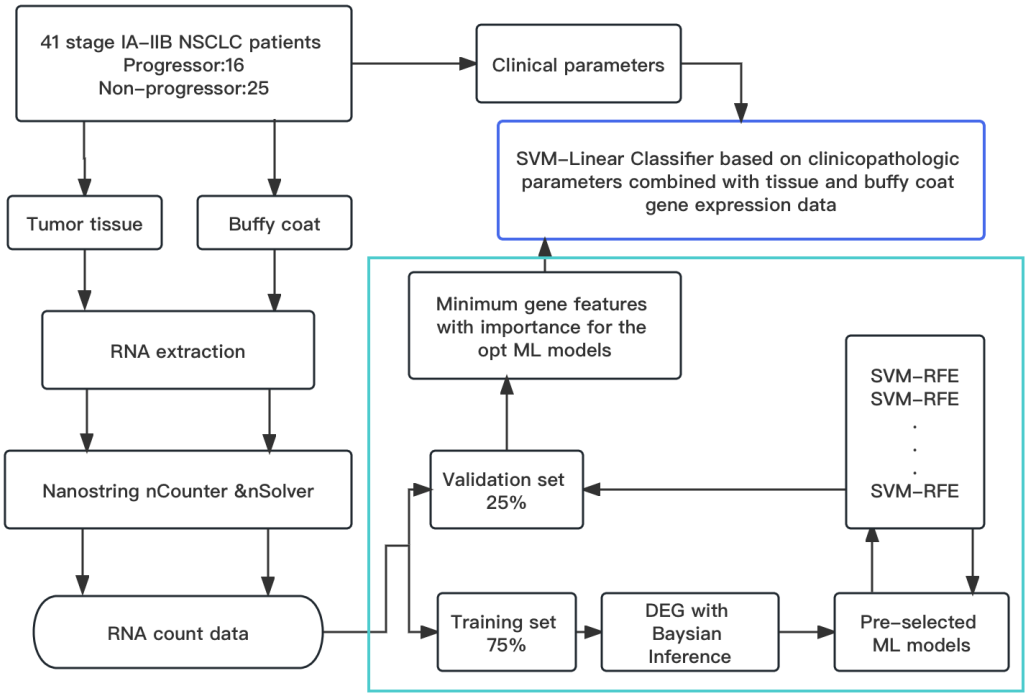
Probe Name	Class Name
PRPF38A	Housekeeping
SAP130	Housekeeping
SDHA	Housekeeping
SF3A3	Housekeeping
TBP	Housekeeping
TLK2	Housekeeping
TMUB2	Housekeeping
TRIM39	Housekeeping
TUBB	Housekeeping
USP39	Housekeeping
ZC3H14	Housekeeping
ZKSCAN5	Housekeeping
ZNF143	Housekeeping
ZNF346	Housekeeping

It is not clear to me that the validation is not involved in the feature selection (as it should not be).

Please make this clearer in figure S1.

Reply 3:

Yes, the validation is involved in the feature selection, we have updated this step in Supplemental Figure 1 (Line 196).



Please indicate HOW you determined the feature-importance ranking; did you look at information gain, Gini index, SHAP, permutation importance, etc.?

Reply 4:

We adopted the CARET package for our ML model development, in the packages a sensitivity analysis was used to measure the effect on the output of any given model when the inputs are varied. The variable importance measure is based on weighted sums of the absolute regression coefficients. For the support vector machine models trained in this study, ROC curve analysis is conducted on each predictor. As we were performing two class problems, a series of cutoffs was applied to the predictor data to predict the class. The sensitivity and specificity are computed for each cutoff and the ROC curve is computed. The trapezoidal rule is used to compute the area under the ROC curve. This area is used as the measure of variable importance.

It is not clear to me how you created the combo-classifier in the context of the pipeline you sketched. It seems that you first performed the SVM-RFE to extract the features, then created the combo-classifier based on the extracted features? I am somewhat skeptical that simply adding expression data and clinical data leads to such a jump in performance, can you explain this?

Reply 5:

Yes, in this study we combined three datasets into our machine learning models, for the gene expression data from nanostring and buffycoat, we trained and validated svm model, then combined the gene expression data with clinical data for the combo-classifier with Multivariate Adaptive Regression Splines method. By including the demographic data and clinical data, the combo model showed improved classification performance.

One low-hanging-fruit addition to this work is the consideration of gene-networks using e.g. string-db.org, i.e. can you discern any pathways with a quick glance?

Reply 6:

Thanks for this valuable suggestion. We have added this content into the manuscript.

Line 47-49:

'Protein-protein network (PPI) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were conducted to identify potential molecular mechanisms underlying tumor progression.'

Line 57-58:

'TNF and IL6 scored the most among the ten hub genes in the PPI network.'

Line 187-193:

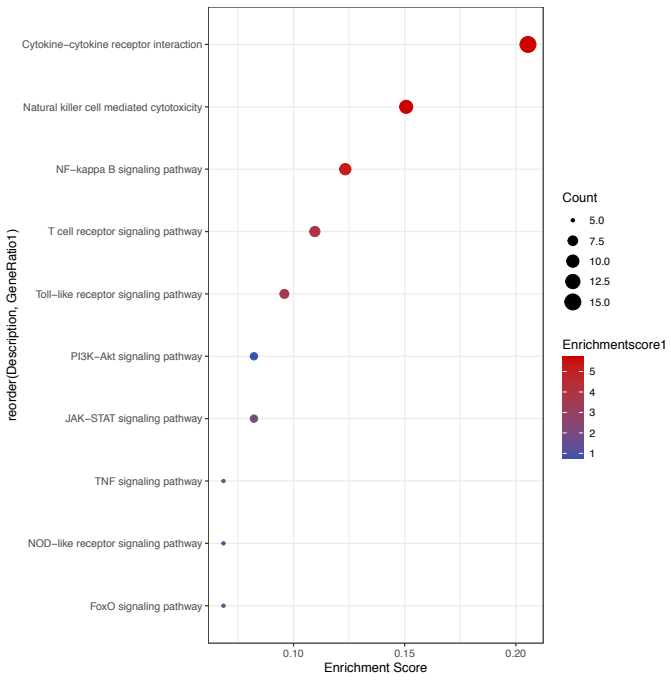
'## Integration of the PPI network and KEGG enrichment analysis of DEGs

The Search Tool for the Retrieval of Interacting Genes version 10.0 (STRING; string-db.org) was used for the exploration of potential DEG interactions at the protein level. Hub genes were identified using the Cytohubba plugin of cytoscape. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was conducted using the R clusterProfiler package to identify DEGs at the biologically functional level. $P<0.05$ was considered to indicate a statistically significant difference.'

Line 271-281

'## Hub genes and KEGG pathway analysis

A total of 92 differential expression genes were screened out by Bayesian inference. In a PPI network containing 89 nodes and 663 edges, ten hub genes (TNF, IL6, CD8A, GZMB, CXCL8, TBX21, PRF1, KLRK1, IRF4 and CD247) were identified. Among them, the score of TNF and IL6 were more than 60 (Supplemental Table 2). KEGG pathway analysis revealed 'Cytokine-cytokine receptor interaction', 'Natural killer cell mediated cytotoxicity', 'NF-kappa B signaling pathway', 'T cell receptor signaling pathway', 'Toll-like receptor signaling pathway', 'PI3K-Akt signaling pathway', 'JAK-STAT signaling pathway', 'TNF signaling pathway ',



'NOD-like receptor signaling pathway' and 'FoxO signaling pathway 'were enriched significantly (Supplemental Figure 2).'

Some details:
"39 T cells, activated natural killer (NK) cells, M0
40 macrophages, and M1 macrophages accounted for a higher proportion in patients with
41 progression (P<0.001, P=0.0089, P<0.001, and P=0.0016, respectively)."
Write explicitly: "T cells... and M1 macrophages were positively associated with"
Reply 7:
Thanks for pointing this out. We have accordingly revised this sentence.
Line 54-57: The proportion of activated natural killer (NK) cells, M0 macrophages, and M1 macrophages were positively associated with progression (P=0.0089, P<0.001, and P=0.0016, respectively). While the proportion of memory resting CD4+ T cells was negatively associated with progression (P<0.001).

There was a significant improvement in accuracy when we fed the linearly combined
46 gene expression data and clinical data into 1 model (AUC of 92.0% in the training set,
47 and 91.7% in the validation set)."
What do you mean with "linearly combined" in this case? I am assuming that you mean "combined in a linear SVM" ?
Reply 8:
We agree with this suggestion. Therefore, we have modified this sentence.
Line 64-66: There was a significant improvement in accuracy when we combined gene expression data and clinical data in a linear SVM model.

Overall:

Very interesting direction of research. I am looking forward for a combination of TIICs, buffy coat and other omics. Pleased by targeted panels instead of WGS which would raise questions about statistical robustness.
Reply 9:
Thanks for your appreciation and suggestion. With these tumor samples and blood samples, we plan to analyze the phenotype of tumor-infiltrating immune cells by mass cytometry next. Then, we'll combine the results of flow cytometry analysis and of the present study to further explore biomarkers which can predict recurrence more accurate.

Reviewer B

1) **First of all, my major concern regarding this study is the small sample size of this study, in particular the 25% validation sample, n=10. The findings from such a small sample are very unstable. The authors obtained satisfactory accuracy parameters of the prediction model, but they deliberately ignored the 95%CIs of these parameters, which should be broader. My second major concern is the predictors including immune gene expression and clinical data were assessed when the outcome of**

recurrence was detected, but for prediction, the potential predictors should precede the outcome of recurrence. The title is unclear, which should clear indicate the development and validation of a prediction model.

Reply 1:

We agree with this comment. Reviewer A also pointed out this question. The small sample size is really an inevitable limitation for this study. That's why we use repeat cross validation in all of our tumor tissue, buffycoat and clinical datasets. We have included this concern in the discussion section (Line 351-352).

For the second concern, since the tumor samples were collected during the operation, and the clinical data were patients' clinical characteristics. No matter tumor recurrence happens or not, both of them won't change. Based on these findings, if we can predict recurrence, then early intervention should be taken.

For the third point, we have changed the title into 'Recurrence Prediction of lung adenocarcinoma using an immune gene expression and clinical data trained and validated SVM classifier'.

2) **Second, the abstract needs some revisions. The background did not explain why g immune gene expression and clinical data and ML algorithm could accurately predict the recurrence and what the clinical significance of this study was. The methods need to describe the inclusion of subjects, the generation of training and validation samples, and collection of clinical variables. The results need to provide the sensitivity and specificity of the predictive model in both the training and validation samples, as well as their 95%CIs. The conclusion needs to be tone down because of the above limitation of this study.**

Reply 2:

Thanks for your valuable suggestions. We have rewritten the abstract.

Line 30-68:

'Background: Immune microenvironment plays a critical role in cancer development, progression, and control. Machine learning algorithm can facilitate the analysis of laboratory results and clinical characteristics of patients for the prediction of cancer recurrence. Early detection and intervention provide the most valuable opportunity for long-term survival in lung cancer relapse. With an aim to evaluate the clinical and genomic prognosticators for lung cancer recurrence, we constructed four machine learning (ML) models and compared their prediction accuracy.

Methods: A total of 41 early-stage lung cancer patients who underwent surgery between June 2007 and October 2014 at Langone Medical Center and had snap-frozen tumor tissue and buffy coat collected at the time of resection were included (with recurrence, n=16; without recurrence, n=25). The Cell-type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) algorithm was used to quantify the fractions of tumor-infiltrating immune cells (TIICs). Protein-protein network (PPI) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were conducted to identify potential molecular mechanisms underlying tumor progression. Each type of data (clinical data, gene expression data of tumor tissue, and buffy coat) were randomly distributed into a training set (75%) and a validation set (25%). Ensemble linear kernel support vector machine (SVM)

ML models were built with both the optimized clinical and genomic features to predict tumor recurrence.

Results: The proportion of activated natural killer (NK) cells, M0 macrophages, and M1 macrophages were positively associated with progression (P=0.0089, P<0.001, and P=0.0016, respectively). While the proportion of memory resting CD4+ T cells was negatively associated with progression (P<0.001). TNF and IL6 scored the most among the ten hub genes in the PPI network. The prediction models based on 12 clinicopathological prognostic factors, expression data of 45 genes from tumors, and 47 genes from buffy coat showed a receiver operating characteristic (ROC) curve area under the curve (AUC) of 62.7% (95%CI: 56.3%-69.1%), 65.4% (95%CI: 59.2%-71.5%), and 59.7% (95%CI: 52.8%-66.5%) in the training set, ROC-AUC of 58.3% (95%CI: 17.9%-98.8%), 83.3% (95%CI: 55.7%-100%), and 75.0% (95%CI: 42.1%-100%) in the validation set, respectively. There was a significant improvement in accuracy when we combined gene expression data and clinical data in a linear SVM model.(AUC of 92.0% in the training set, and 91.7% in the validation set).

Conclusions: Using ML algorithm, immune gene expression data from tumor tissue and buffy coat may help improve the accuracy of lung cancer recurrence prediction.'

3) **Third, in the introduction of the main text, a brief review on known biomarkers and clinical factors that can predict or be associated with reoccurrence is needed. The authors need to comment their limitations and explain why a combination of immune gene expression and clinical data can potentially accurately predict the reoccurrence. The further question is the strength of ML algorithm.**

Reply 3:

Thanks for your comments. We have revised the introduction part and added some references which focus on lung cancer biomarkers.

Line 95-101:' Yu et al.(15) construsted a nomogram model based on smoking, solid nodules, mucinous lung adenocarcinoma and micropapillary component. The internal and external validation C-indexes of the nomogram were 0.822 (95% CI: 0.751–0.891) and 0.812, respectively. Genetic predisposition also involved in tumor recurrence. Single nucleotide variants of MSH5, MMP9 and CYP2D6 were found significantly associated with early-stage LUAD presenting with GGNs (16).'

Line 106-108: 'Bacterial biomarkers also played a role in predicting the survival of lung cancer patients. The relative abundances of bacteria were significantly different between the recurrence group and non-recurrence group (19).'

The small sample size is really an inevitable limitation for this study. Only 41 samples were included, which may not be sufficient to draw a firm conclusion. We have included this into our limitation paragraph (Line 399-402). Besides, too many genes were included in the combo-classifier, which may make it inconvenient and costly in clinical application. Thus, future development of a simpler combination of genes which does not sacrifice accuracy would be preferable.

The clinical model based mostly on the staging system, which is too broad to predict prognosis precisely and help guide treatment. While more and more evidence show that gene-related biomarkers improve prediction accuracy. Expression of genes changes

throughout the development of cancer, it provides more information than clinical features. That's why we combine gene expression data with clinical data into one model. And there is a big jump in the performance of the individual classifiers and the combo-classifier.
As a popular mathematical tool, ML can improve the accuracy of cancer prediction by 15–20%. Using ML algorithm, we repeated cross validation in all of our tumor tissue, buffycoat and clinical datasets to remedy the small sample size issue to some extent.

4) **Fourth, in the methodology of the main text, the authors need to clearly describe the clinical research design, sample size estimation, the calculation of AUC, sensitivity, and specificity, as well as the 95%CIs of these accuracy parameters.**
Reply 4:
Thanks for your suggestion. We have presented the study design in Figure S1.
Line 196: 'The entire research design is shown in Supplemental Figure 1.'
And we also add the following sentences into the methods part. Line 216-221: "ROC curve was plotted using the pROC package in R, AUC, sensitivity and specificity was computed. 95% CIs for sensitivity, specificity, and AUC were computed using bootstrapping techniques with the boot package in R."

5) **Finally, several related papers should be reviewed and cited in this study: 1. Jeong WG, Choi H, Chae KJ, Kim J. Prognosis and recurrence patterns in patients with early stage lung cancer: a multi-state model approach. Transl Lung Cancer Res 2022;11(7):1279-1291. doi: 10.21037/tlcr-22-148. 2. Fu R, Zhang JT, Chen RR, Li H, Tai ZX, Lin HX, Su J, Chu XP, Zhang C, Qiu ZB, Chen ZH, Tang WF, Dong S, Yang XN, Zhang GQ, Zhao GP, Wu YL, Zhong WZ. Identification of heritable rare variants associated with early-stage lung adenocarcinoma risk. Transl Lung Cancer Res 2022;11(4):509-522. doi: 10.21037/tlcr-21-789. 3. Yu S, You C, Yan R, Chen H, Chen C, Xu S, Gonzalez M, Chen R, Kang M, Chen S. Establishment and validation of a nomogram model for predicting postoperative recurrence-free survival in stage IA3 lung adenocarcinoma: a retrospective cohort study. Transl Lung Cancer Res 2022;11(11):2275-2288. doi: 10.21037/tlcr-22-776.**
Reply 5:
Thanks for your suggestion. We have added the three reference into the manuscript.
Line 95-98:'Yu et al.(15) construsted a nomogram model based on smoking, solid nodules, mucinous lung adenocarcinoma and micropapillary component. The internal and external validation C-indexes of the nomogram were 0.822 (95% CI: 0.751–0.891) and 0.812, respectively.'
Line 98-101:'Genetic predisposition also involved in tumor recurrence. Single nucleotide variants of MSH5, MMP9 and CYP2D6 were found significantly associated with early-stage LUAD presenting with GGNs (16).'
Line 314-315:'Factors associated with recurrence include histological, clinical, and population-based characteristics (27,28).'