Peer Review File
Article information: https://dx.doi.org/10.21037/jtd-22-1312

Reviewer A

General Comments for Authors:
In this paper, the authors proposed a decision tree model based on CT findings to predict EGFR mutation status in synchronous multiple primary lung cancers (SMPLC). They found CT-based decision tree model is a simple tool to predict the status of EGFR mutation that may be considered for treatment decision-making. However, my main concern focuses on the statistical methods part. The description of the statistical methods is not clear. The authors need to clarify their methods carefully before the reviewers can check whether the method is applied appropriately and whether the results are valid. More detailed suggestions/comments are as follows:

1. In line 169 on page 8, the authors mentioned they applied the Lasso method to select potential predictors of EGFR mutation among all clinicopathological features and CT signs. However, two lambdas are selected. I think this part is very confusing. Could the authors clarify whether they applied the lasso method to select clinicopathological features and CT signs in two separate regression models? If the authors use two separate models, could they explain why they did not include all clinicopathological features and CT signs in one model and use the lasso method to select? It is more common to use lasso methods to select all variables simultaneously. If the authors use one model to select all clinicopathological features and CT signs, could the authors explain whether two lambdas are provided?

Reply: We are extremely grateful to the reviewer for pointing out this problem. We indeed select clinicopathological features and CT signs in two separate regression models based on the clinical workflow. The clinicopathological features and imaging features were two important elements in clinical diagnosis and treatment, especially in predicting the EGFR mutation. So we tend to select the features separately, to select significant variables. **we have modified our text as advised** (see Page 7, line 160-162): "Least absolute shrinkage and selection operator (LASSO) regression was used to select the clinicopathological features and CT signs in two independent regression models related to EGFR mutation to develop a DTA.".

2. In line 173 on page 8, the author used a decision tree model to predict EGFR mutation. Different covariates, splitting rules, and stopping rules in the decision tree model may lead to different results. However, the author did not provide these details. Could the authors provide the covariates included in the decision tree at the beginning, the splitting rules, and the stopping rules?

Reply: The features selected by LASSO regression were input to build decision tree model by the classification and regression trees (CART) algorithm. A possible decision or an action result in binary groups. Two child nodes are then generated from a parent node and this tree-growing methodology leads to the best split based on the splitting criterion. During this course, every child node will become a parent node when it splits. The decision-making process stops when no contribution exists in the further branching. Gini impurity was used for impurity measurement.

We have modified our text as advised (see Page 7, line163-171):"Using the factors identified on Lasso regression analyses, conditional inference tree analyses were performed to construct a DTA (JMP pro V.14.3, SAS Institute, Cary, NC, USA). The features selected by LASSO regression was input to build decision tree model by the classification and regression trees (CART) algorithm. A possible decision or an action results in binary groups. Two child nodes are then generated from a parent node and this tree-growing methodology leads to the best split based on the splitting criterion. During this course, every child node will become a parent node when it splits. The decision-making process stops when no contribution exists in the further branching. Gini impurity was used for impurity measurement. "

3. In line 179 on page 8, the author exhibited the calibration curves to evaluate the agreement between outcomes predicted by the DTA model and the real-world outcomes. The results in this part are confusing. Firstly, in the right subfigure of Figure 5, the authors compared the predicted and actual probabilities for EGFR mutant strains. Could the author clarify how the actual probabilities for EGFR mutant strains can be obtained? Secondly, it seems that the authors applied the model to the whole dataset to get predicted probabilities and compared them to the real outcomes. In statistics, if the entire dataset has been used to fit the model, comparing the whole dataset's prediction and real outcomes may be unfair because of the overfitting issues. The more standard method is to use cross-validation to avoid overfitting issues. Could the author clarify whether the predicted probabilities are calculated from the whole dataset used to fit the model?

Reply: Thanks for your comments. The data was grouped into small groups after bootstrapping and the actual probability of EGFR mutations was calculated within each group, compared to the mean predicted probability. Smoothing techniques can be used to estimate the observed probabilities of the outcome in relation to the predicted probabilities. Furthermore, the internal validation was reached by bootstrapping methods. And Steyerberg EW et al. found the bootstrapping method was recommended to estimate internal validity of a predictive logistic regression model.

4. In the "Further evaluation of the CT-based DTA method" section, the authors said they built a logistic regression model with the novel CT-based multi-parameter decision tree algorithm model and other clinicopathological characteristics derived from the LASSO regression analysis. This part is unclear to me. Could the authors explain how the novel CT-based multi-parameter decision tree algorithm model can be included in the logistical regression? Did the authors use the predicted probabilities

from the novel CT-based multi-parameter decision tree algorithm model as the covariate in the logistic regression? If so, the model includes an estimated variable as one covariate. In this case, the p-values from the standard logistic regression functions in R, such as the glm() function, are invalid. Could the authors specify which function in R they used for logistic regression?

Reply: According to the novel DTA model, the parameter was calculated and included in the logistical regression for each patients, which was similar to risk score instead of predicted probabilities in the model. And the glm() function was used for logistic regression.

Minor concerns:

1. Figures 3 and 4 are not clear to read. For example, the covariate names of figures in the left column in Figure 3 are not readable. The numbers of the importance of Figure 4 are not legible. The authors should make the graphs more readable, such as increasing the font size of the text.

Reply: Thanks for your kindly suggestions, **we have modified the figure3**. Could Figure 4 be added as supplementary material, because the information of the figure was too much, increasing the font size of the text was invalid.

2. The caption of Figure 3 needs to be modified. The captions of the figures should be relevant to the content of the figure. In Figure 3, the authors do not need to re-explain how the lasso method works in variable selection but need to explain what the four subfigures represent.

Reply: Thanks for the reviewer, **we have modified the figure legends of figure 4 as advised** (see Page 17, line410-413)

"LASSO coefficient profiles of the clinicopathological features (A) and CT signs (C), respectively. Dotted vertical lines are drawn at the optimal values using the minimum criteria and the 1 standard error of the minimum criteria (the 1-SE criteria) using the optimal log(λ)=-2.5 and -3.5, respectively (B, D)."

3. In Figure 5, could authors explain "apparent," "bias-corrected," and "ideal" in the legend of the lines?

Reply: The original Figure 4 was added as supplemental figure, so the original Figure 5 was corrected as "Figure 4" in the revised manuscript. The "apparent" line represented the best-fit model, and "bias-corrected" was derived from training dataset by bootstrap approach, and "ideal" line represented ideal model, which actual probability was equal to predicted probability.

We modified the figure legends of figure 4 as advised (see Page 17, line416-419):

"(B) The 45° dashed line illustrates the ideal prediction that the actual probability is equal to the predicted probability. The plot represents the accuracy of the best-fit model ("Apparent") and the bootstrap model ("Bias-corrected") for predicting the EGFR mutations."

Reviewer B

In this manuscript, Luo and colleagues report on the clinical utility of CT-based decision tree model for predicting EGFR mutation status in synchronous multiple primary lung cancers. This manuscript is of some interest. The results are clearly presented and well discussed. This simple method could help clinicians choose an appropriate therapy in patients with high risk of complications.

My minor comment is as follows:
1) It is often challenging to distinguish SMPLC from intrapulmonary metastases, solely based on CT characteristics. The authors should add some discussion about the difference between them.
Reply: Thanks for your comments, **We have modified our text as advised** (see Page 10-11, line241-247):
"In patients with multiple pulmonary sites of involvement, distinguishing between multiple primary lung cancers (MPLCs) and intrapulmonary metastasis (IPM) is critical for developing a therapeutic strategy. Suh et al. applied one algorithm based on comprehensive information on clinical and imaging variables that allows differentiation between MPLCs and IPMs. Furthermore, predicting the status of EGFR mutation in MPLCs might facilitate personalized precision treatment of these patients (21)"
[1] *Suh YJ, Lee HJ, Sung P, Yoen H, Kim S, Han S, Park S, Hong JH, Kim H, Lim J, Kim H, Yoon SH, Jeon YK, Kim YT. A Novel Algorithm to Differentiate Between Multiple Primary Lung Cancers and Intrapulmonary Metastasis in Multiple Lung Cancers With Multiple Pulmonary Sites of Involvement. J Thorac Oncol. 2020 Feb;15(2):203-215. doi: 10.1016/j.jtho.2019.09.221. Epub 2019 Oct 18. PMID: 31634666.*

2) Please describe the details of EGFR mutations, if possible.
Reply: Thanks for your comments. In our study, we developed a people-oriented and CT-based decision tree model for predicting EGFR mutation status in SMPLCs. Based on patient-level analysis, the subjects were divided into two groups, which were EGFR mutation (n=53) group and EGFR wild-type (n=32) group.

Reviewer C

In this study, the authors aim to create a CT-based decision tree model for predicting EGFR mutation status in synchronous multiple primary lung cancers. Although the results looked promising, some major points should be addressed as follows:
1. Sample size is too small (about 85 patients) and this amount may affect the significance of their findings.

Reply: Thanks for your comment. The incidence of MPLCs in patients has been reported as 0.2% to 8%[1]. The diagnostic criteria of SMPLC was strict that even require molecular genetics and NGS test. In our study, we excluded SMPLC patients as follows: (a) thin-slice chest CT was not available; (b) not undergoing surgical resections; (c) lymphadenectomy not performed; (d) preoperative treatment prior to surgery, such as radiation therapy or chemotherapy; (e) concurrent or previous other malignancy; (f) testing results for EGFR mutation status not available. Thus, 85 SMPLCs were enrolled in this study. The sample size was similar to some recent studies on SMPLC.

2. The authors should have external validation data to evaluate the performance of model on unseen data.

Reply: The major limitation of the present study is the lack of external validation. We have tried to collect data from other centers, but very little data meets our inclusion criteria. Further external validation in a large independent population is still required. Nevertheless, in our subgroup analysis, this simple CT-based decision tree model had good discrimination for outcome prediction, and could be easily performed in clinics.

3. It is unclear how the authors selected decision tree as their optimal model although recently there are a lot of advanced models may be better than decision tree. Thus, the authors should have a comprehensive comparison to proof their choice.

Reply: The features selected by LASSO regression was input to build decision tree model by the classification and regression trees (CART) algorithm.

The method has several advantages compared with these other machine learning algorithms, including: (1) several comparative studies have shown higher accuracy and versatility for DTA than other machine learning methods [1,2]; and (2) the DTA model structure is less complex than other methods [3], which facilitates model interpretation and reduces the need for model optimization, the DTA model has stronger explanatory power, provides a visual representation of the decision-making process, allowing for easy guidance and application; (3) previous studies verified the decision tree model can be reliably employed to mine large sets of input variables, in relatively small samples [4].

[1] Zhou X, Xu J, Zhao Y. Machine learning methods for anticipating the psychological distress in patients with Alzheimer's disease. Australasian Physical & Engineering Sciences in Medicine 2006, 29:303–309.

[2] Tighe P, Laduzenski S, Edwards D, Ellis N, Boezaart AP, Aygtug H.Use of machine learning theory to predict the need for femoral nerve block following ACL repair. Pain Med 2011, 12:1566–1575

[3] Freund Y , Mason L . The Alternating Decision Tree Learning Algorithm. Morgan Kaufmann, 2002.

[4] Rousseau S, Polachek IS, Frenkel TI. A machine learning approach to identifying pregnant women's risk for persistent post-traumatic stress following childbirth. J Affect Disord. 2022 Jan 1;296:136-149. doi: 10.1016/j.jad.2021.09.014. Epub 2021 Sep 22.

PMID: 34601301.

4.The authors should compare the predictive performance to previously published works on the same problem/data.

Reply: To our knowledge, this study is the first to build a CT-based decision tree model for predicting EGFR mutation status in SMPLC patients. Compared with the study published in 2020, For prediction of EGFR mutations in multiple pulmonary adenocarcinoma , the author used logistic regression analysis and receiver operating characteristic curve (ROC) analysis yielded area under the curve (AUC) values of 0.647 and 0.712 for clinical-only or combined CT features, respectively. In our study, ROC analysis reached an AUC of 0.854, was significantly superior to this previous study. **We have modified our text as advised** (see Page 13, line292-297): "To the best of our knowledge, this is the first study to build a CT-DTA model for predicting EGFR mutation status in SMPLC patients, and ROC analysis reached an AUC of 0.854. The results of our study are significantly superior to the previous study by Han et al. (15). Moreover, the DTA tree model has strong explanatory power and provides a visual representation of the decision-making process for easy guidance and application."

5. Besides current CT features, the authors may add CT-based radiomics features since they have been proven efficient in such kind of study.

Reply: Thanks for your suggestion. When planning a radiomics study, a key consideration is to determine the availability of sufficient data to support the development of a radiomics signature. For example, using the "one-third" criteria and a 10-feature model, at least 133 samples are required [1]. Now the sample size in our study was too small to conduct a radiomics study, we will enrolled more cases in our next study.

[1] Shur JD, Doran SJ, Kumar S, Ap Dafydd D, Downey K, O'Connor JPB, Papanikolaou N, Messiou C, Koh DM, Orton MR. Radiomics in Oncology: A Practical Guide. Radiographics. 2021 Oct;41(6):1717-1732. doi: 10.1148/rg.2021210037. PMID: 34597235; PMCID: PMC8501897.

6. The model may be improved using deep learning models.

Reply: Thanks for your suggestion. The sample size in our study was too small to conduct a deep learning study, we will enrolled more cases in our future study.

7.ROC or AUC is well-known and has been used in previous biomedical studies i.e. PMID: 36166351, PMID: 35767281. Thus, the authors are suggested to refer to more works in this description to attract a broader readership.

Reply: Thank you for your suggestion, the reference were added.

8.English writing should be improved.

Reply: Thanks for your suggestion, The wordings of the main text and/or figures/tables will be checked by a native English-speaking expert who is majoring in this field.

9. Quality of figures should be improved.

Reply: Thanks for your suggestion, The wordings of the main text and/or figures/tables will be checked by a native English-speaking expert who is majoring in this field.

10. Some technical aspects and essential insights of the proposed method are not described in detail.

Reply: Thanks for your kindly comment. **We have modified our text as advised** (see Page 7, line163-171):"Using the factors identified in LASSO regression analyses, conditional inference tree analyses were performed to construct a decision tree algorithm (DTA , JMP pro V.14.3, SAS Institute, Cary, NC, USA). The features selected by LASSO regression were input to build a DTA model using the classification and regression trees (CART) algorithm. A putative decision or an action resulted in binary groups. Then, two child nodes were generated from a parent node, and this tree-growing methodology leads to the best split based on the splitting criterion. During this splitting, every child node becomes a parent node. The decision-making process stops when no contribution exists in the further branching. Gini impurity was used for impurity measurement."

11. The review of related work is not sufficiently thorough and not sufficiently specific.

Reply: Thanks for your kindly comment. To our knowledge, little research has been done on this subject, and our study is the first to build a CT-based decision tree model for predicting EGFR mutation status in SMPLC patients.