Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-24-38

Reviewer A

Comment 1: It is unclear why the authors are focused particularly on bone mets. Why not evaluate for risk of any metastatic disease? Brain mets in particular would be very important to know for purposes of radiation therapy. While the authors mention the importance of bone mets and their impact on survival, they don't make clear why they are excluding other sites for metastatic disease.

Reply 1: Thank you. Lung adenocarcinoma is the largest subtype of lung cancer. Compared with other subtypes, lung adenocarcinoma has a higher incidence of bone metastasis. About 20%-50% of lung adenocarcinoma patients will develop bone metastasis(1, 2), and about 30% of lung adenocarcinoma patients will develop brain metastasis(3). Patients with bone metastasis have worse prognosis and lower survival rate than those with brain metastasis(4), and the median survival time is only 6-10 months. In addition, there is a lack of effective treatment after the occurrence of bone metastases, and there is no clear evidence of the long-term efficacy of traditional antineoplastic therapy for patients with bone metastases. In this condition, it is particularly important to develop a risk stratification tool for bone metastasis in lung adenocarcinoma patients to identify patients at high risk of bone metastasis as early as possible and provide timely intervention. We made changes in the Introductory section of the revised manuscript accordingly.

Changes in the text: we have modified our text as advised (see Page 3-4 line 70-86)

1. Decroisette C, Monnet I, Berard H, et al. Epidemiology and treatment costs of bone metastases from lung cancer: a French prospective, observational, multicenter study (GFPC 0601). J Thorac Oncol. 2011; 6(3):576-82.
2. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. Clin Cancer Res. 2006; 12(20 Pt 2):6243s-9s.
3. Cagney DN, Martin AM, Catalano PJ, et al. Incidence and prognosis of patients with brain metastases at diagnosis of systemic malignancy: a population-based study. Neuro Oncol. 2017; 19(11):1511-21.
4. Riihimaki M, Hemminki A, Fallah M, et al. Metastatic sites and survival in lung cancer. Lung Cancer. 2014; 86(1):78-84.

Comment 2: How were the clinical factors chosen for inclusion? Is there some data that shows that these particular factors are predictors of lung cancer prognosis? There should be a citation here to relevant manuscripts.

Reply 2: Thank you for your advice. The selection of clinical factors was based on previous studies of risk factors for BM in lung cancer, and we have added relevant citation information to the manuscript. We made changes in the Methods section of the revised manuscript accordingly.

Changes in the text: we have modified our text as advised (see Page 5 line 129-134)

Comment 3: The segmentation description is a bit confusing. Were all tumors segmented by a single radiologist, then an additional 60 were segmented by 2 radiologists? This should be clarified.

Reply 3: Thank you very much for your correction. The images were initially segmented by Radiologist 1, following which 60 patients' CT images were randomly chosen from the entire dataset and re-segmented by Radiologist 2. We made changes in the Methods section of the revised manuscript accordingly.
Changes in the text: we have modified our text as advised (see Page 6 line177-179)

Comment 4: One of the biggest issues is that the radiomics model, as written in the methods, appears to have been created after using that same data to choose the best features for feature reduction. This will automatically bias the model to a high performance since it has already "seen" the data. The authors should consider holding out a small segment of their data set to perform feature reduction, then building the model on the remainder of their data. That way, the model is never exposed to the data on which feature reduction was performed.
Reply 4: Thank you for your valuable advice. Regarding your concern about possible label leakage during model training and validation, label leakage is a common problem in machine learning and can lead to overestimation of model performance, that is, overfitting. In order to avoid label leakage and overfitting, we carefully considered when designing experiments. First, we applied cross validation in Lasso algorithm. Secondly, based on the type IIa research standard of the TRIPOD specification, the research subjects were divided into two independent data sets: training and validation. The independent samples of the validation set were used to verify whether the model was overfitting. Finally, we did not use algorithms such as smote to expand or reduce the sample size, which may lead to overfitting of the model. It is the best strategy to design three sets of training, testing and validation. However, due to the insufficient sample size in this study, we only designed training and validation sets. We will continue to increase the sample size in the future to establish a more scientific and rigorous model through three sets, which will be elaborated in the limitations of the study. Thank you again for your advice.
Changes in the text: we have modified our text as advised (see Page 7 line 192-196, Page 11 line 341-346)

Comment 5: There is significant class imbalance here. There are almost 4 times as many patients without bone mets as with. Therefore, if the model were to predict all patients were not going to have bone mets, it would be right almost 72% of the time! In creating model such as this, methods can be used, for example undersampling or SMOTE, to better balance the classes.
Reply 5: Thank you for your suggestions and your concerns. Sample imbalance will lead to insufficient classification learning for small samples, which will lead to poor classification diagnosis for small samples. However, in real-world medical research, we often encounter, or in most cases, sample imbalance, which is mostly caused by the epidemiology of the disease or the characteristics of the disease itself. For example, only 1/5 patients in this study have bone metastasis, which is the real-world metastasis rate. In machine learning research, although oversampling of small sample classification (represented by the SMOTE classical algorithm) can improve this problem in some cases, it is not applicable in many cases, because these algorithms to expand samples are easy to lead to overfitting, and in many cases cannot really improve the performance

of the model. In this study, for the bone metastasis population, the sensitivity, specificity and accuracy of our model were 0.750, 0.913 and 0.883, respectively. The performance of our model was good in the validation cohort and the training cohort. Secondly, after the application of SMOTE, the AUCs of the training cohort and the validation cohort were 0.959 and 0.785, respectively. The performance of the model in the validation cohort was not as good as that in the training cohort, and overfitting occurred. For the above reasons, we did not use SMOTE for sample expansion, we discussed it in the Discussion section of the revised manuscript.
Changes in the text: We have revised our text as suggested (see Page 9 line 263-283)

Comment 6: There are no 95% confidence intervals provided for the AUCs, nor has a comparison of the values to generate a p-value been performed. Looking at the results, it seems highly likely that the performance of each method falls within the 95% CI of the other methods. It therefore seems difficult to believe that the nomogram out-performs the other models. It would be important to include a 95% CI here for the reader to fully assess each model's performance.
Reply 6: Thank you very much for your advice. We provide 95% confidence intervals provided for the AUCs in Table 2, and in response to your suggestion, we have modified Figure 4 so that these results are presented more clearly to the reader. In addition, we performed Delong tests on the performance of the three prediction models. The results showed that the AUC values of the radiomics model and the combined model were statistically different (p=0.033) in the training cohort, and the AUC values of the other models were not statistically different, detailed results are provided in the Supplementary material.
Changes in the text: We have revised our figure and text as suggested (see Page 7 line 213-214, Page 8-9 line 251-255)

Reviewer B

Comment 1: However, I would like to suggest further exploration and discussion on two important aspects: data imbalance and excluded samples. Addressing these topics could enhance the comprehensiveness of your analysis and provide readers with a deeper understanding of the challenges and considerations involved in data analysis and machine learning.
Reply 1: Thank you for your advice. For the data imbalance, sample imbalance will lead to insufficient classification learning for small samples, which will lead to poor classification diagnosis for small samples. However, in real-world medical research, we often encounter, or in most cases, sample imbalance, which is mostly caused by the epidemiology of the disease or the characteristics of the disease itself. For example, only 1/5 patients in this study have bone metastasis, which is the real-world metastasis rate. In machine learning research, although oversampling of small sample classification (represented by the SMOTE classical algorithm) can improve this problem in some cases, it is not applicable in many cases, because these algorithms to expand samples are easy to lead to overfitting, and in many cases cannot really improve the performance of the model. In this study, for the bone metastasis population, the sensitivity, specificity and accuracy of our model were 0.750, 0.913 and 0.883, respectively. The performance of our model was good in the validation cohort and the training cohort. Secondly, after the application of SMOTE,

the AUCs of the training cohort and the validation cohort were 0.959 and 0.785, respectively. The performance of the model in the validation cohort was not as good as that in the training cohort, and overfitting occurred. So, we did not use SMOTE for sample expansion, we discussed it in the Discussion section of the revised manuscript.

For the exclusion sample, which was a cohort study, it is worth mentioning that patients with synchronous bone metastases at baseline (n=279) were excluded in order to accurately quantify the association with risk factors and to better focus on the process of bone metastasis. We have revised the manuscript accordingly.

Thanks again for your advice!

Changes in the text: We have revised our text as suggested (see Page 5 line123-126, Page 9 line 263-283)