Peer Review File

Article Information: https://dx.doi.org/10.21037/tlcr-22-248

Reviewer A

This work is of big scientific value and in my opinion, it brings daily radiomic application closer to clinical implementation.

Comment 1

Please verify:

Line 38: missing space before No evidence…

Line 99: one space to much after (21,22)

Reply 1: We would like to thank the reviewer for the feedback and for the attentive revision of our work, we have corrected the spacing as indicated.

Changes in the text 1: line 38 and line 99, tracked within the text of the re-submitted manuscript

Comment 2

The title is clear and informative.

The Abstract provides clear summary of the study background, methods, results, and conclusions.

Keywords are appropriate.

Reply 2: We thank the Reviewer for their positive comments on the above-indicated sections of the manuscript.

Changes in the text 2: None

Comment 3

In Introduction the Authors describe background concerning medical imaging, why they choose NSCLC CT for this study and importance of radiomics implementation to the clinics.

I find the Author's objectives essential:

Quantify the feature/volume dependency across multiple preprocessing methodologies and volume groups (high vs low).

Assess whether these variations have an impact on survival model performance.

Serve as a hypothesis-generating work.

Reply 3: We really appreciate the Reviewer's comment on the rationale and objectives of our work.

Changes in the text 3: None

Comment 4

Materials and Methods section is divided in 6 parts which are all very well written with all analysis's details. The datasets used were retrieved from publicly available repository, which gives a broad opportunity for other scientists to proceed with these studies and check the Authors hypothesis which is one of their objectives. Models trained with and without volume, baseline model, low-volume, high-volume model have been compared.

In my opinion, the analyzes are very thoughtful and valuable, also for me personally as a

radiologist cooperating with other professionals (thoracic surgeons, oncologists etc.).

Reply 4: We thank the Reviewer for their positive comments on the above-indicated sections of the manuscript.

Changes in the text 4: None

Comment 5

Also, results section is written with all the details about the analyses and survival models making the paper understandable for the lay person.

Reply 5: Thank you for your positive feedback

Changes in the text 5: None

Comment 6

In Discussion the Authors identified and listed limitations of the study but also the strengths.
I agree that in this paper it has been shown that the use of different preprocessing methods has a potentially relevant impact on feature/volume correlation, the features dependence from tumor volume is a critical issue in radiomic studies.

Reply 6: Thank you for your kind comment

Changes in the text 6: None

Comment 7

The figures and tables are clear and informative.
I think this work brings daily radiomic application closer to clinical implementation.

Reply 7: We are delighted of the positive comments of the Reviewer about our work.

Changes in the text 7: None

Reviewer B

Comment 1

The explorative study is well written with an interesting design.
However, I would recommend submitting the article to a more suitable journal with a stronger bioinformatic background and audience

Reply 1: We would like to thank the Reviewer for the overall positive feedback on our study. We had initially considered a more bioinformatics- oriented journal for the submission, but after discussion among all Authors, we have agreed that TLCR would have been a more suitable target to share our manuscript with a larger research community. Specifically, we strongly believe that, after 10-years of radiomics, also clinicians are required to achieve a more solid methodological background, and that wider accessibility of such works through translational journals could be a great contribution towards the achievement of this goal.

Changes in the text 1: None

Major limitations:

Comment 2

-The study only includes data of the publicly available repository curated by TCIA. Clinical variables are quite rare in the database and may have a significant impact such as treatment modality and tumor biological features.

Reply 2: Thank you for the comment. We are aware that the lack of recognized clinical features is a shared limitation of our work, and of all manuscripts using such public datasets. As pointed out in the Discussion section, the absence of other oncological outcomes (e.g. cancer-specific survival, loco-regional progression-free survival) can easily be indicated as a weakness. In this, we completely agree with the Reviewer's observation. However, this was intended as a hypothesis-generating, proof-of-concept study, and we believe that the use of a large, public, well-known dataset could be beneficial in encouraging further analysis on preprocessing and feature-volume correlation.
Changes in the text 2: None

Comment 3
-In addition, an external validation of the findings would highly strength the manuscript.

Reply 3: We agree that external validation is critical for determining the robustness of the observations, as we wrote in the Discussion section, last paragraph: "Additionally, these models currently lack validation on external datasets, which would help to achieve higher robustness". This is a potential direction for further research.
Changes in the text 3: None

Reviewer C
Comment 0
The manuscript assesses the impact of image filters on the volume-dependence and prognostic value of features in the open-source NSCLC-Radiomics data. The article is well written, and the author displays a detailed understanding about the wider literature in lung cancer radiomics research. I have some suggestions to improve the manuscript below:

Reply 0: Thank you for the comment, below follows a point-wise reply to the Reviewer's comments.
Changes in the text 0: None

Comment 1
1. I feel the purpose of the study is not clear throughout, in the introduction the author states the intent was to provide a 'methodological framework' to identify reproducible and informative features (which has been previously done by Traverso et al, https://doi.org/10.1016/j.ejmp.2020.02.010). The conclusions of the study, however, do not make strong suggestions on a framework with advice to have 'informative data-sets', 'further modelling techniques' and 'external validation' which are not methods to correct for volume-confounding (which should be the focus here).

Reply 1: Thank you for the comment. We have slightly reformulated both the aim of the study and the conclusion of our work to address the issues raised by the Reviewer hoping that the new version now complies with their remarks.
Changes in the text 1: Introduction section lines 39-44, Conclusion section lines 321-335.

Comment 1a

a. Remove the sentences on developing a 'methodological framework' to make it clear the aim is to assess the impact of image filters on volume dependency and survival modelling for different size lesions in the introduction.

Reply 1a: Thank you for the comment, we have modified the manuscript accordingly**.**

Changes in the text 1a: Introduction section line 43-44, tracked within the text of the re-submitted manuscript.

Comment 1b

b. Ensure the impact of filters on both volume-dependency and survival is considered in the conclusion and abstract.

Reply 1b: The conclusion section has been rephrased so as to better reflect the core aim of the study, clearly indicating the conclusions from both the volume analysis and survival analysis.

Changes in the text 1: Conclusion section line 306-322, tracked within the text of the re-submitted manuscript

Comment 1c

c. Can you make any future recommendations relevant to image pre-processing in the conclusion?

Reply 1c: Explicit recommendations are now included in the conclusion.

Changes in the text 1c: Conclusion section line 306-322, tracked within the text of the re-submitted manuscript

Comment 2

2. As stated in the title, the article aims to assess the impact of "image pre-processing" on the volume-dependence of radiomic features, however, only image filters are assessed ignoring the potential impact of voxel-size resampling and bin-discretisation which are other important aspects of pre-processing that could influence the volume dependence. To improve on this, I would make the following suggestions:

Reply 2: Thank you for the feedback, below follows a point-wise reply to the Reviewer's comments

Changes in the text 2: none

Comment 2a

a. Report if there was any resampling and what bin width/bin number was used.

Reply 2a: No resampling was used since the maximum deviation from the median spacing was 0.16 mm (only present in one patient). The bin width was set to 25 (the PyRadiomics default).

Changes in the text 2a: this information has been added to section 2.3 (line 72), tracked within the text file.

Comment 2b

b. I think it is a large limitation if the role of resampling/binning was not considered which should be discussed and tested if possible. Especially because filters used will impact whether resampling can be done and what bin discretisation should be applied (https://arxiv.org/pdf/2006.05470.pdf).

Reply 2b: The title and the aim of the study have been slightly reformulated to better reflect the fact that resampling and binning are not considered. We agree that this is an important aspect that should be tested if possible. It may however be more suitable to investigate these consequences in more heterogeneous datasets and in multiple modalities, which would be a rather tall order for this manuscript.

Changes in the text: Title has been changed, tracked within the text of the re-submitted manuscript

Comment 2c

c. Change the title and wording from 'image pre-processing' to focus on image filters if other parameters are not tested.

Reply 2c: Thank you for the suggestion, we have changed the title accordingly.

Changes in the text 2c: Title modified according to the Reviewer's suggestion.

Comment 3

3. I do not agree with the way the statistical model comparison is done in Figure 3. The C-index represents models built on different sub-groups of the same cohort patients (all, low volume, high volume) where patient demographics, event rate and sample size will also differ. I think the test should be 'clinical + volume + radiomics' vs 'clinical + volume' models in each case using likelihood statistics (nested model comparison). From this result you can determine in which sub-group does radiomics improve most upon the clinical model.

Reply 3: Thank you for pointing out this correction. We have included the suggested analysis in the new manuscript (however, we opted for Wilcoxon signed-rank test, since we have multiple repeated pair-wise observations, and the CatBoost model is not a likelihood estimator)

Changes in the text 3: Materials and Methods section line 156-159, Results section line 212-216, Discussion section line 295-297.

Comment 4

4. Did the author consider volume dependence in the low-volume group vs the high-volume group? I.e., if filters have more of an impact at small or low volumes.

Reply 4: We did not investigate this specific relationship, but we do agree that it is an interesting direction. If we buy into the observation that noise seems to dominate over feature values the lower the ROI volume is, one would expect the volume dependence to reflect this in the low volume group.

Changes in the text 4: None

Comment 5

5. Page 8, line 346 'did not lead to improvement' – should be re-worded as it made the stratification worse.

Reply 5: Thank you for the comment, we have re-worded the sentence accordingly.
Changes in the text 5: Discussion section (line 248), tracked within the text of the re-submitted manuscript

Reviewer D
Comment 1
This paper reflects a large amount of work looking at the cross-comparison of CT preprocessing filters and radiomic feature extraction to tumor volume size and the impact of survival prediction model performance. The work illustrating the impact of pre-processing on feature/volume correlation is interesting and informative. It is advantageous that publicly available data (Lung 1) and tools (pyradiomics and python model implementation) were used in the study. Although accessibility is slightly lessened due to the decision to have a single Radiologist adjust the GTV supplied with the public dataset. There are major concerns with the conclusions drawn from the results presented.

Reply 1: We would like to thank you the Reviewer for the feedback and the comments on our work. We agree that the editing of the publicly available contours from the TCIA dataset may slightly impair the reproducibility of our work in other centers. However, we considered that, given the nature of the study, the exclusion of pathological lymph nodes from the volume of interest would be a tool to overcome possible sources of variability from radiomic features inhomogeneities between primary and nodal volumes. We think that a reasonable solution to account for the above-mentioned considerations would be to make the segmentation dataset available to all interested researchers upon request to the Corresponding Author.
Changes in the text 1: Discussion section (line 296-298).

Comment 2
- The first sentence of the abstract conclusion mentions feature/volume dependence should be adequately managed to limit overfitting. However, this is not clearly addressed in your study. The box and whisker plots for the prediction models (Figure 2) reflect a large range in performance across cross-validation runs – which appears to be approximately consistent with and without volume. No external validation was performed (Lung 2 data?).

Reply 2: The observation that the performances appear consistent/similar is correct (comparing models with volume vs without volume). At the same time, it is also evident that the inclusion of radiomic features does in fact add independent informative content. This can be inferred from the same plot, but by comparing “clinical” with “all”. To better support and clarify this conclusion/observation, we have modified the structure of the conclusion section and results section (including a new plot - Figure3 - specifically illustrating this).
The overfitting/colinearity issue may be viewed as a corollary from the results in Figure 1a, which in turn (often) renders the differences in Figures 2 a & b insignificant. In other words, we should not be too surprised that the exclusion of volume has little impact when we know that many radiomic features convey similar information as volume (as evident from Figure 1a). We hope that the updated conclusion and results reflect these considerations in a more conspicuous manner.

With regards to the Lung2 dataset, we agree that it is a very promising and interesting direction for further validation. At present, however, some information relevant to the purpose of our work is lacking, such as the clinical T- and N- stages. More importantly, the population of the Lung2 dataset is characterized exclusively by early-stage NSCLC patients treated with surgery, which rises major concerns about its comparability with the Lung1 dataset. Nevertheless, we understand the importance of providing further testing and validation of our results, and we are currently working towards this aim using a comparable cohort from our Institution.

Changes in the text 2: Conclusion section (line 306-322). Modified the conclusion section in both the abstract and main text. Added a plot illustrating the added value of radiomics in the results.

Comment 3

- The second conclusion sentence in the abstract is not informative nor helpful. Prognostic models are developed with the intention to assist with difficult clinical cases (most commonly the earlier stage, smaller lesions). This sentence phrasing almost encourages cohort selection bias.

Reply 3: Thank you for the comment, the new conclusion should better reflect the core messages we are trying to get across. We should have solved the Reviewer's concerns in the updated version.

Changes in the text 3: Conclusion section (line 306-322), tracked within the text of the re-submitted manuscript.

Comment 4

- A major goal of this work was related to tumor volume bias, however, no details are provided regarding the volumes of the low/high cohorts. Please provide further detail (mean+/- SD) for the low and high cohorts.

Reply 4: Thank you for the suggestion, we have now added the median value and the related IQR in the manuscript.

Changes in the text 4: Materials and Methods section (line 151-152), tracked within the text of the re-submitted manuscript.

Comment 5

- The division of the tumor into low and high-volume categories is based solely on the midpoint of the data publicly provided (and likely doesn't reflect a clear clinically meaning separation of patients but does capture other biases – ie likely the large volume cohort had more late-stage subjects). This limitation needs to be addressed

Reply 5: This is indeed a potential limitation. The primary reason this separation was chosen was that we did not agree upon a robust cutoff point that made sense from a clinical perspective. For instance, the dataset is strongly imbalanced with respect to early vs locally advanced stages. Given the exploratory nature of the work, we believe the median value to be

a good compromise.
Changes in the text 5: none

Comment 6
- There is very little variation in the moderate c-index performance (or SD) across all models and conditions for survival prediction (Table 1) which raises the question if all models have issues with stability. Testing on a validation cohort (Lung 2) would be beneficial (and needed if overfitting is to be addressed).
Reply 6: We do not believe model stability should be a major cause for concern in this context since the performance measures are collected from varying training sets. However, analysis on a held-out validation set (e.g. Lung 2) would be a welcomed addition to the study, as noted above. At present, however, we feel that including this analysis might divert the study from the primary questions we want to address, as better discussed above (comment 2).
Changes in the text: None.

Comment 7
- The results in figure 3 showing significant differences in all models looking at the performance in low/high tumor volume cohorts – do not indicate that the studies process for removal of feature dependence from tumor volume was effective. This does not align with the conclusions as stated.

Reply 7: Our primary reason for removing features with high volume correlation was to exclude trivial relationships from the analysis. Even so, excluding stronger volume relationships from the feature data would not necessarily undermine the conclusions that can be drawn from the figure. In any case, we hope that the reformulated conclusions should be more in line with our observations.
Changes in the text 5: none