Peer Review File

Article information: https://dx.doi.org/10.21037/tlcr-23-84

Review comments-Reviewer A

1) First, the title did not indicate the survival outcome to be predicted by the model to be developed and the comparisons between deep transfer learning-based vs. Cox regression-based.

   Reply 1:

   We included overall survival in lung cancer as the survival outcome to be predicted in the title.

   However, the comparisons between the deep transfer learning-based vs. Cox regression-based are the subjective statistic to evaluate the performance, but not the main focus of our study. Out study promotes a new way to apply machine learning based modeling in medical field.

   Moreover, development and/or validation of the model and outcome are required to be included in the title according to TRIPOD checklist, but not the comparisons. We would like to add only the outcome but not the comparisons to keep the title simple.

   Changes in the text:

   Title was changed to "Development and validation of a deep transfer learning-based multivariable survival model to predict overall survival in lung cancer" in Page 1.

2) Second, the abstract is not adequate and needs to be revised. The background did not indicate the clinical needs for the model based on deep learning and transfer learning and the potential strengths of such model. The methods did not describe the generation of training and validation samples, including the interval validation sample, the assessment of variables used in the predictive model, and the survival outcomes to be predicted. The results need to describe the comparisons of AUC values between the two models. The conclusion needs more detailed comments for the clinical implications of the findings.

   Reply 2:

   We revised the background of the abstract to emphasize the clinical needs of our study.

   We described how we developed primary model using SEER database, internally validated with SEER sample and externally validated with GYFY database. Due to the word count limit of the abstract, we believe that details regarding generation of the samples and assessment of the variable should not be included in the abstract.

   Changes in the text: Page 2 Line 23-26; Page 3 Line 14-17.

3) Third, the introduction of the main text needs to review available algorithms used in machine learning that have been used to predict the survival in cancer with particular focus on their predictive accuracy. Please also analyze the potential strengths and increased accuracy of the model based deep learning and transfer learning and explain why it can improve the predictive accuracy.

   Reply 3:

   We included some available machine learning models for lung cancer in the introduction.

   Changes in the text: Page 4 Line 17-30.

4) Fourth, the methodology of the main text cannot be so simple and so inadequate like this. The authors need to describe the clinical research design, describe the SEER data sources, describe the variables extracted from the SEER dataset, the survival outcomes to be predicted, i.e., 3-year or 5-year, and the generation of the training and validation samples. In statistics, please describe how the new and cos models were established. Please describe the threshold values of AUC for a good predictive model, as well as the calculation of sensitivity and specificity, which are also important accuracy parameters. Detailed algorithms, procedures, and formulas can be provided in the supplementary file. This part should report the overall frame work of the analysis.

   Reply 4:

   More information regarding statistical analysis was added into the Method section.

   More details regarding research design was described in the text and the Figure 1.

   SEER data resources, variables extraction, detailed algorithms and formula were already included in the supplementary.

   Changes in the text: Page 5 Line 20 – Page 6 Line 4.

Review comments-Reviewer B

Summary:

The paper presents a new survival model for lung cancer that integrates deep learning and transfer learning. While the author's motivation for improving upon the Cox model and the study's aims and scope are clear and well-justified, the presentation of the methodology lacks clarity and reproducibility. The paper's organization could be improved for better flow and readability, and there are numerous grammatical and word choice errors that impede the reader's comprehension of the text.

My major concern about the paper is that the authors claim to introduce a novel approach to survival prediction by utilizing a non-linear model in place of the traditional Cox's model, while also incorporating Transfer Learning (TL). However, DeepSurv, the model they employed, has been in use for cancer survival predictions since 2018, and several articles have already utilized

it. While the authors employed transfer learning in conjunction with the Cox model, they did so in a rather standard manner. Consequently, the novelty of their approach may not be as significant as suggested in the paper.

Recommendations:

Major comments:

1. The study claims to introduce a novel approach to survival prediction by utilizing a non-linear model in place of the traditional Cox's model, while also incorporating Transfer Learning (TL). To accomplish this, the authors employed DeepSurv (Katzman et al., 2018), an established model for survival analysis. However, it is worth noting that DeepSurv has been in use for cancer survival predictions since 2018, and several articles have already utilized it. The authors assert that their unique contribution lies in the application of transfer learning. However, their approach to transfer learning appears to be fairly standard and lacking in innovation, with no particularly novel methods employed.

Reply 1: You are certainly correct that DeepSurv, as a matter of fact, machine learning was employed to great effect in the production of the cancer survival models; further, it is true that transfer learning was also used previously to the same effect. What we proposed in our study that is novel is the co-application of both DeepSurv with the addition of transfer learning that would make our model more robust in dealing with the future within the context of lung cancers.
Change in the text: None.

2. It would be beneficial to compare the proposed non-linear model with other existing models in the field of survival prediction, beyond just the Cox model. Merely comparing it to the Cox model does not provide enough insights, especially considering that DeepSurv, even without transfer learning, has already demonstrated better performance in predicting survival than the Cox model. This has been previously shown in papers such as Katzman et al., 2018 and Kim et al., 2019. It is unclear what the authors' contribution is to using DeepSurv out-of-the-box. Therefore, it would be helpful if they compared their proposed model to other existing models, not just the Cox model.

Katzman, Jared L., Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network." BMC medical research methodology 18, no. 1 (2018): 1-12.

Kim, Dong Wook, Sanghoon Lee, Sunmo Kwon, Woong Nam, In-Ho Cha, and Hyung Jun Kim. "Deep learning-based survival prediction of oral cancer patients." Scientific reports 9, no. 1 (2019): 1-10.

Reply 2: You are certainly correct in mentioning that DeepSurv, even without transfer learning, has already demonstrated better performance in predicting survival than the Cox model. What we are proposing, which we also consider to be novel, is the addition of Transfer learning onto the DeepSurv allows our model to be more robust. And it allows our model to be remodeled by other researchers using the transfer learning and explore other prognostic factors, which is new.
Change in the text: None

3. The paper is not well-organized or easy to follow. Some terms are mentioned before they are properly introduced and explained. As a result, the reader has to "jump" between sections because the authors explain the advantages and results of the model without first addressing their new model's name, what kind of model it is, and how it is different from previous ones.

Reply 3: Please let us know more specifically which part of the article confused you most, so we can better edit our article.
Change in the text: None.

Minor comments:

1. The text lacks clarity and concision, and is riddled with errors. While the study's goals and motivation are clearly stated, the descriptions of the model and its validation are too broad and vague. Moreover, there are numerous grammatical and word choice errors that impede the reader's comprehension of the text.

Reply 1: We did have our article reviewed by an editor who has English as their native language. We also have re-written and re-checked our paper for this re-submission. If you have noticed any further grammatical mistakes or awkward word choices, we apologize and ask that you specifically point them out so that we can fix them.
Change in the text: Page 2, Line 22-28

2. "Interests" might not be the best word choice for "Many deep learning-based survival models are emerging in different disease interests, but those integrating deep learning and transfer learning are rare."

Reply 2: Edited as requested
Change in the text: Page 2, Line 22-23

3. Page 3, line 27:
"Given the relatively miserable outcome of lung cancer," change miserable for "poor" .

Reply 3: Edited as requested
Change in the text: Page 3, Line 29

4. Page 3, line 30:
The acronym “TNM” was not previously introduced.

Reply 4: Edited as requested
Change in the text: Page 4, Line 1

5. Page 4, line 29: “The outcome” instead of just “outcome”.

Reply 5: Edited
Change in the text: Page 5, Line 9

6. On Page 4, line 30, by “the model”, are you referring to your model? It is not entirely clear from the context.

Reply 6: Yes. Edited to “our model”
Change in the text: Page 5, Line 10

7. Page 5, line 1: The “DeepSurv” model was not previously introduced. In addition, the appropriate citation is missing for most of the paper and is only introduced and referenced on Page 7, line 12.

Reply 7: Added the reference
Change in the text: Page 5, Line 12

8. The median ages for the cohorts are more than a decade apart. This could skew the results, but it is not listed as one of the limitations.

Reply 8: We used the SEER database, the biggest lung cancer database that includes 601,480 patients, as the training cohort to explore the coefficients of prognostic factors. After pre-training, our model could provide fixed coefficients of the 18 variables that were closest to the real coefficients among the global population, which enabled our model to perform accurate predictions in different populations globally. And the pretrained model still gave good performance in GYFY cohort in our external validation, which represents the stable performance of the model in populations with different characteristics.
Change in the text: None

9. DeepSurv and the Cox model are compared only based on the Concordance indexes, perhaps it would be useful to complement this comparison using a different metric. You mentioned DeepSurv had superior accuracy and AI certainty, but the scores for these metrics were not included along the Concordance indexes.

Reply 9: Concordance indexes were chosen for its familiarity and ease of data presentation. We judged that our audience would be best served with only one data point that encapsulates the standard performance measure for our model assessment in survival analysis
Change in the text: None

10. On page 6:
The train/test splits for each dataset were 80/20 for SEER and 70/30 for GYFY. Why did you use two different kinds of train/test split, is there a special reason for it?

Reply 10: The different split was chosen because otherwise we would not have enough sample population for study comparison.
Change in the text: none

11. It is not clear in the main text if you are using the term "certainty" as the definition you give in the Appendix for "AI certainty" or as a synonym for "accuracy". I would suggest clarifying in the main text to which term you are referring when you mention "certainty" the first time. (Page 7, lines 20 and 21, for example)

Reply 11: It is AI certainty throughout the main text. Edited as needed in main text.
Change in the text: Page 2 Line 12; Page 3 Line 5,9,16; Page 5 Line 15; Page 7 Line 15; Page 8 Line 22; Page 11 Line 10

Review comments-Reviewer C

This paper innovatively integrates deep learning and transfer learning to achieve higher generalization through fine-tuning. In the case of missing data, it is a more stable and rigorous model than the traditional one.

Suggestions:
Proportion of different races in SEER and GYFY cohorts are relatively unbalanced, in which white race takes up to 82%. Meanwhile, the Asian race is missing, which may cause race bias.

Reply: We used the SEER database, the biggest lung cancer database that includes 601,480 patients, as the training cohort to explore the coefficients of prognostic factors. After pre-training, our model could provide fixed coefficients of the 18 variables that were closest to the real coefficients among the global population, which enabled our model to perform accurate predictions in different populations globally. While white population is 82% other population still has a sample size of 37,541. And the pretrained model still gave good performance in GYFY cohort in our external validation, which represents the stable performance of the model in populations with different characteristics.

Change in the text: None