Peer Review File

Article Information: https://dx.doi.org/10.21037/cco-23-85

Response to Reviewer A

Comment 1: Overall the results are important. However, it is very difficult for these results to be applicable for the common reader who has no familiarity with machine learning. The types of algorithms must be better explained in detailed so that a broader audience may be able to take broader learning points away.

Reply 1: Thank you for your comments sincerely. We have modified our text as advised, we have added "The ST is a machine learning method that constructs a tree structure by splitting nodes by maximizing survival differences until all terminal nodes containing only the minimum number of unique events(31, 32). Both RSF and GBM are combined of a large number of survival trees. RSF uses the bootstrap method to extract sub samples from the original samples to construct a survival tree, averaging the cumulative risk function of each survival tree and ultimately obtaining the total cumulative risk function(33). GBM is a machine learning method based on gradient descent boosting. The fundamentals of GBM is training a new survival tree according to the negative gradient information of the loss function based on the current survival tree, and combining the trained newborn survival tree with the existing survival tree(34)." in the second paragraph of Introduction Part (see Page 6-7, line 85-94).

Changes in the text: We have added "The ST is a machine learning method that constructs a tree structure by splitting nodes by maximizing survival differences until all terminal nodes containing only the minimum number of unique events(31, 32). Both RSF and GBM are combined of a large number of survival trees. RSF uses the bootstrap method to extract sub samples from the original samples to construct a survival tree, averaging the cumulative risk function of each survival tree and ultimately obtaining the total cumulative risk function(33). GBM is a machine learning method based on gradient descent boosting. The fundamentals of GBM is training a new survival tree according to the negative gradient information of the loss function based on the current survival tree, and combining the trained newborn survival tree with the existing survival tree(34)." in the second paragraph of Introduction Part (see Page 6-7, line 85-94).

Comment 2: In addition, the authors should focus on highlighting a more practical utilization of their results as additional future directions within their discussion. For example, how can this technology be utilized practically to help patients?

Reply 2: We have modified our text as advised. We have added "The proposed nomogram can be used to calculate the three-year and five-year CSS of gastric cancer patients based on the clinical information. It may be utilized practically to help clinicians to obtain individualized survival

prediction and provide better treatment allocation." in the last paragraph of the Discussion Part (see Page 22-23, line 426-429).

Changes in the text: We have added "The proposed nomogram can be used to calculate the three-year and five-year CSS of gastric cancer patients based on the clinical information. It may be utilized practically to help clinicians to obtain individualized survival prediction and provide better treatment allocation." in the last paragraph of the Discussion Part (see Page 22-23, line 426-429).

Comment 3: In terms of predictors for their analysis, the authors should clarify what type of stage (AJCC)- Clinical or Pathologic. I am assuming they are referring to AJCC Clinical Stage, 8th edition.

Reply 3: In this study, the type of stage (AJCC)- Clinical or Pathologic were referring to AJCC Clinical Stage, 8th edition. We have added "The predictors of TNM stage, T stage, N stage, M stage were referring to AJCC Clinical Stage, 8th edition." in the first paragraph of "2.2.Predictors and outcomes" Part (see Page 9, line 136-137).

Changes in the text: We have added "The predictors of TNM stage, T stage, N stage, M stage were referring to AJCC Clinical Stage, 8th edition." in the first paragraph of "2.2.Predictors and outcomes" Part (see Page 9, line 136-137).

Comment 4: It may be helpful to include neoadjuvant vs adjuvant chemotherapy and radiation separately as predictors. This may be important in terms of additional stratification and relevance.

Reply 4: It' s a great suggestion for "include neoadjuvant vs adjuvant chemotherapy and radiation separately as predictors". However, the variables related to chemotherapy in the SEER database only include whether chemotherapy has been implemented, and don't involve specific chemotherapy methods. In addition, there are many missing cases with specific chemotherapy methods in our development databset. We are terribly sorry to haven't included neoadjuvant vs adjuvant chemotherapy and radiation separately as predictors.

Changes in the text: No change.

Comment 5: The learning data set and external test data set have very different survival times. Is this appropriate to use as a development set given different outcomes. This seems to be a very different cohort. The authors must include more on this in the discussion. Do they think this is a reflection of treatment or tumor biology?

Reply 5: The different survival times of the two datasets, may be related to treatment or tumor biology. In our study, we think may be due to Hebei Province was a high-risk area for gastric cancer and had relatively advanced gastric cancer diagnosis and treatment technologies. In addition, it's may due to the different follow-up time. The maximum follow-up duration were 5 years and 9 years for development dataset and external dataset, respectively. Thus, we have added "In addition,

it's may be related to treatment, tumor biology or follow-up time. Further exploration is still needed." in the end of second paragraph of Discussion Part (see Page 19, line 343-344).

Changes in the text: We have added "In addition, it's may be related to treatment, tumor biology or follow-up time. Further exploration is still needed." in the end of second paragraph of Discussion Part (see Page 19, line 343-344).

Comment 6: For figures 5 and 6 as well as the supplementary data using the models, they data is not presented with appropriate transitions/explanations within the text. It is very difficult for the reader to understand the significance of the models and how this relates to the multivariable models. This needs to be spelled out better for the reader, particularly for those who do not have high familiarity with machine learning. The discussion does not fully explain the significance of these findings either.

Reply 6: Thank you for your comments sincerely. The Figure 5, Figure S4 and Figure S7 depicted the calibration curves of different models. The calibration curve,like C-index and AUC, is one of evaluation indicators for the model. The calibration curve is used to evaluate the consistency between the predicted values and the actual observed values of models. In order to help the reader better understand the calibration curve, we have added an explanation "The 45-degree straight gray line of calibration curve represents the perfect match between the observed (y-axis) and predicted (x-axis) survival probabilities. A closer distance between two curves indicates higher consistency." in the seventh paragraph of Methods Part (see Page 10, line 157-160). In order to presented with appropriate transitions/explanations within the text for the Figure S5 and Figure 6, and make readers easier to understand the significance of the models and how this relates to the multivariable models, we have added "From above analysis, we can achieve that the predictive performance of Cox and RSF were superior to that of ST and GBM, and the predictive performance of Cox was similar to that of RSF. Due to the fact that the current application of RSF wasn' t as simple as Cox, in order to better apply the model to practice, we visualized the Cox' s results by drawing a forest plot and constructed a nomogram for clinicians to predict patients' survival." in "3.4.Outcome of Cox" Section of Results Part (see Page 15, line 262-267). In order to make reader easier to understand, we have given a fully explain of these findings as reviewer's advised. We have added "ST splitting nodes by maximizing survival differences among nodes using log-rank testing. However, the prediction error is large, resulting in low prediction accuracy(31, 32). Both RSF and GBM are combined of a large number of survival trees. The fundamentals of GBM is training a new survival tree according to the negative gradient information of the loss function based on the current survival tree, and combining the trained newborn survival tree with the existing survival tree(34). In this study, The C-index and AUC of GBM are similar to that of Cox or RSF. However, the consistency of calibration curve of GBM performs poorer, which means it needs to be improved. RSF uses the bootstrap method to o extract sub samples from the original samples to construct a survival tree, averaging the cumulative risk function of each survival tree

and ultimately obtaining the total cumulative risk function(33).” in the third paragraph of Discussion Part (see Page 19, line 350-361).

Changes in the text: We have added an explanation “The 45-degree straight gray line of calibration curve represents the perfect match between the observed (y-axis) and predicted (x-axis) survival probabilities. A closer distance between two curves indicates higher consistency.” in the seventh paragraph of Methods Part (see Page 10, line 157-160). We have added “From above analysis, we can achieve that the predictive performance of Cox and RSF were superior to that of ST and GBM, and the predictive performance of Cox was similar to that of RSF. Due to the fact that the current application of RSF wasn' t as simple as Cox, in order to better apply the model to practice, we visualized the Cox' s results by drawing a forest plot and constructed a nomogram for clinicians to predict patients' survival." in “3.4.Outcome of Cox" Section of Results Part (see Page 15, line 262-267). We have added “ST splitting nodes by maximizing survival differences among nodes using log-rank testing. However, the prediction error is large, resulting in low prediction accuracy(31, 32). Both RSF and GBM are combined of a large number of survival trees. The fundamentals of GBM is training a new survival tree according to the negative gradient information of the loss function based on the current survival tree, and combining the trained newborn survival tree with the existing survival tree(34). In this study, The C-index and AUC of GBM are similar to that of Cox or RSF. However, the consistency of calibration curve of GBM performs poorer, which means it needs to be improved. RSF uses the bootstrap method to o extract sub samples from the original samples to construct a survival tree, averaging the cumulative risk function of each survival tree and ultimately obtaining the total cumulative risk function(33).” in the third paragraph of Discussion Part (see Page 19, line 350-361).

Comment 7: During introduction, the authors write tumor write infiltration, this should be tumor size.

Reply 7: In the introduction, the tumor infiltration represent the T stage of gastric cancer. In the AJCC TNM 7th, the T stage means the tumor infiltration and tumor size. Thus, we have replaced the “tumor infiltration” with “tumor infiltration, tumor size” (see Page 5, line 61).

Changes in the text: We have replaced the “tumor infiltration” with “tumor infiltration, tumor size” (see Page 5, line 61).

Comment 8: The authors should use the same terminology throughout the paper (training set vs development dataset.

Reply 8: We have modified our text as advised. We have replaced “set” with “dataset” (see Page 2, line 32; Page 10, line 171).

Changes in the text: We have modified our text as advised. We have replaced “set” with “dataset” (see Page 2, line 32; Page 10, line 171).

Response to Reviewer B

Comment 1: One of the important parameters influencing survival in gastric cancer patients is systemic treatment, such as chemotherapy. However, this parameter cannot be included in the model due to not meeting the proportional hazards (PH) assumption. This may lead to potential inaccuracies in predicting survival. Therefore, it is advisable to use a statistical analysis approach that is more flexible and fully parametric, beyond the Cox model, which is a semi-parametric model when used for building the survival prediction model.

Reply 1: Thank you for your suggestions for our manuscript sincerely. It's a great suggestion to use a statistical analysis approach that is more flexible and fully parametric, beyond the Cox model. As you can see, we are struggling to explore better survival analysis models than Cox. However, due to our limited technology, we only mastered three tree based machine learning methods: ST, RSF, and GBM. In future research, we will attempt to incorporate more models into our analysis, hoping to find better survival prediction models. We are terribly sorry to haven' t use a statistical analysis approach that is more flexible and fully parametric.

Changes in the text: No change.

Comment 2: ECOG performance status and nutritional status are clinical prognostic factors commonly used in clinical practice. However, this study did not incorporate them into the prognostic model, as they may not be fully comprehensive for practical medical application.

Reply 2: It' s a great suggestion for incorporating ECOG performance status and nutritional status into the prognostic model. However, there are few cases having ECOG performance status and nutritional status in our development databset. And, there aren't the two variables in the external test dataset. If we include them, it will result in the loss of development dataset and a large number of cases in the external test dataset. We are terribly sorry to haven' t included ECOG performance status and nutritional status as predictors. We have added "ECOG performance(60)" in the last but one paragraph of Discussion Part (see Page 22, line 419).

Changes in the text: We have added "ECOG performance(60)" in the last but one paragraph of Discussion Part (see Page 22, line 419).

Comment 3: In fact, the differences in data between the development dataset used to create the model and the external test dataset used for external validation are not considered a limitation of the research. Such differences can actually serve as a valuable tool for effective external validation.

Reply 3: Thank you for your suggestions for our discussion sincerely. We have modified our text as advised. We have added “Such differences between development dataset and external test dataset can serve as a valuable tool for effective external validation.” in the fifth paragraph of Discussion Part (see Page 21, line 400-401).

Changes in the text: We have added “Such differences between development dataset and external test dataset can serve as a valuable tool for effective external validation.” in the fifth paragraph of Discussion Part (see Page 21, line 400-401).