Reviewer A

This manuscript tried to identify the risk factor for the respiratory failure necessitating HFNC among the elder patients (70 years old or older) who underwent thoracic surgery. They also assessed chest X ray grading scale can be helpful to identify the risk.

My main concern is how they set the primary outcome. The primary outcome of this study was the incidence of respiratory failure necessitating HFNC, which is very questionable whether this is meaningful in the post operative care. Comparing to re-intubation, we can easily put HFNC. And detecting the potential needs for HFNC may not be very important. Also, the criteria of upgrading to HFNC might not be very vigorous because it is easily reversible. Therefore, I am not quite sure whether this study will be very valuable for the reader and thoracic surgery community.

Second concern is RALE score and Brixia score are grading scale for the non-surgical patients. Thus, these may not be validated for the use of post op patients especially for somebody who underwent lung resection.

Comment 1: My main concern is how they set the primary outcome. The primary outcome of this study was the incidence of respiratory failure necessitating HFNC, which is very questionable whether this is meaningful in the post operative care. Comparing to re-intubation, we can easily put HFNC. And detecting the potential needs for HFNC might not be very important. Also, the criteria of upgrading to HFNC might not be very vigorous because it is easily reversible. Therefore, I am not quite sure whether this study will be very valuable for the reader and thoracic surgery community.

Reply 1: Thank you for your comment. The significance of our study can be delineated as follows: 1. Despite routine postoperative chest X-rays for patients undergoing thoracic surgery within 24 hours, we observed an underutilization of the information derived from these X-rays, with infrequent instances of quantifying lung injury based on them. 2. Our findings reveal that over half of the patients requiring HFNC therapy experienced this need beyond the initial 24 hours postoperatively. Notably, our chest X-ray reassessment, conducted within 24 hours, allows our model to effectively anticipate the necessity for escalated oxygen therapy without imposing an additional burden. 3. In developing countries, COT remains the primary postoperative respiratory support. Premature adoption of HFNC not only places an economic burden on patients but also garners attention from medical insurance. Our study demonstrates that leveraging information from postoperative chest X-rays can enhance predictions of COT failure. Therefore, the identification of risk factors and predictors for AHRF and HFNC requirement is crucial for early detection and intervention.

The criteria of upgrading to HFNC were based on the current guidelines and evidence for oxygen therapy in AHRF. We defined AHRF as PaO_2/FiO_2 100-300 mmHg or SpO_2 < 92%, with increased respiratory rate, heart rate, and exclusion of hypoxemia from other complications(https://doi.org/10.1001/jama.2012.5669).

Comment 2: Second concern is RALE score and Brixia score are grading scale for the non-surgical patients. Thus, these may not be validated for the use of post op patients especially for somebody who underwent lung resection.

Reply 2: Thank you for your comment. We agree that the RALE scoring systems were originally developed and validated for ARDS. However, we have learned that some studies have applied the RALE score to the evaluation of patients after lung, esophagus, and heart surgery. For example, Lizhen Xuan et al (https://doi.org/10.21037/jtd-23-822). used the RALE score to assess the severity of bilateral lung injury in patients after lung cancer surgery. This article was published in the JTD journal in October 2023. Xiaoliang Leng et al (https://doi.org/10.1053/j.semtcvs.2021.03.033). applied the RALE score to the postoperative evaluation of esophageal cancer patients and proposed a new method to identify acute lung injury (ALI). Karim Mostafa et al (https://doi.org/10.3390/jcm12186043). considered the RALE scoring system to be a practical tool that can objectively assess unilateral pulmonary edema (UPE) on chest X-ray images after mitral valve surgery (MVS). Therefore, we believe that these scoring systems are also applicable and useful for assessing postoperative pulmonary complications in elderly patients after thoracic surgery. Moreover, we are the first to demonstrate that the RALE score can be used to predict the risk of COT failure.

As for the Brixia score, we also agree that it was designed specifically for COVID-19 patients, but there is no literature to date that indicates that the Brixia score is not suitable for assessing the severity of pneumonia in postoperative patients. Therefore, we only used the Brixia score for postoperative patients in an exploratory manner. We found that although the Brixia score seemed to predict COT failure well in the univariate analysis, it was excluded in the multivariate analysis. We think this may be because the Brixia score and the RALE score are strongly correlated, so the multivariate analysis only selected the RALE score, which had a better predictive effect. However, whether the Brixia score can really be used for the prediction and evaluation of postoperative patients still needs to be verified by multicenter, large-sample prospective studies.

Changes in the text: We have modfied our title as advised (line 125-126).

Minor concerns are

1. The authors should elaborate p value higher than p>0.001. There are inconsistency exhibit p-value.

Reply 1: Thank you for your comment. We apologize for the confusion caused by the inconsistency in reporting p-values. We have revised the manuscript to report p-values consistently as follows: p-values less than 0.001 are reported as p<0.001, and p-values between 0.001 and 0.05 are reported as p=0.0xx, where xx is the exact value.

Changes in the text: We have modified our text (see line 238, 241, 242, 246).

2. Why the median score of Brixia and RALE of COT failure group has increased after the propensity score matching. The table shows n = 19 in both matched and non-matched analysis.

Reply 2: Thank you for your comment. Tables 1 and 4 present the median and

interquartile range of Brixia and RALE scores for the COT success group and COT failure group before and after propensity score matching. The matching process aimed to reduce the covariate imbalance between the two groups, resulting in a more uniform and comparable sample. For the COT failure group (n=19), propensity score matching was conducted with the COT success group (n=223), matching factors such as age, gender, BMI, etc., at a 1:2 ratio. This resulted in 34 cases being successfully matched from the COT success group, meaning that each COT failure case was paired with two COT success cases having similar propensity scores within the specified range. Consequently, some lower-scoring COT success cases were excluded from the matched sample, leading to an increased median for the COT success group, while the median for the COT failure group remained unchanged. Due to the limitations in the total sample size, only 34 patients were successfully matched out of the 223 COT success patients.

3. The authors need to re-classify the surgical procedure into Esophagectomy (McKewon, Ivor-Lewis), lobectomy, Segmentectomy, wedge resecton etc. The current classification is just surgical approach.

Reply 3: Thank you for your comment. We also considered grouping by surgical procedure as you suggested when we designed the experiment. However, we found that due to the limited sample size and the diversity of mediastinal surgery, grouping by surgical procedure would result in too few cases in the other group (in the manuscript we defined all surgeries except lung and esophagus surgery as other), lacking statistical power. Therefore, we decided to group by surgical approach, which we think is more relevant and appropriate for our study. We will change the term "surgical procedure" to "surgical approach" in the manuscript to make it more accurate. Thank you for pointing out this issue.

Changes in the text: We have changed the term "surgical procedure" to "surgical approach" in the manuscript (including figures and tables).

Reviewer B

I think this study is conducting detailed.

The novelty may be the use of Brixia score and RALE score.

However, since X-ray evaluation is routinely performed postoperatively, I think it is natural that it correlates with the need for oxygen therapy.

Postoperatively X-ray is required, but SpO2 is sufficient to consider the need for oxygen therapy. Therefore, the usefulness of this study is unclear.

Reply: Thank you for your comment and feedback. We appreciate your recognition of the novelty of our study in using the Brixia score and RALE score to evaluate chest X-ray findings in elderly patients after thoracic surgery. We would like to address your concern about the usefulness of our study as follows:

We acknowledge that to some extent, SpO_2 levels can determine oxygen therapy strategies. However, based on the cases we collected, half of the patients upgraded to HFNC treatment within 24 hours postoperatively, even though their SpO_2 levels were

normal during this period. This suggests a potential lag in the feedback provided by SpO$_2$. Nevertheless, through the analysis of chest X-rays within 24 hours postoperatively and consideration of other risk factors, we can proactively adjust oxygen therapy strategies and enable early intervention. Therefore, we believe that chest X-ray scores can complement SpO$_2$ in guiding oxygen therapy and clinical decision-making in postoperative patients. This is especially important in resource-limited settings, where COT may be more cost-effective than HFNC, but also carries the risk of delayed recognition and treatment of hypoxemic respiratory failure.

Reviewer C

This is a very well written paper and a certainly very interesting topic. However, the background of the project and the exectution are to my opinion a little difficult to interpret. Two scoring systems, which are used for other purposes were used to quantify infiltrates in the lung on the first postop. day and this was used to develop a predictive score based on observations in less than 20 patients? It is well known that around 10% of our patients will develop postop. respiratory failure and we very well know that thoracotomy, surgical time etc. are risk factors for respiratory failure. And it is pretty sure that if the authors checked factors like preop lung function and exercise capacity, these factors would also make it to the score system. I am finding it difficult to accept the clinical use of an additional scoring system which is made up of already known risk factors and lung infiltrates.

Comment 1: This is a very well written paper and a certainly very interesting topic. However, the background of the project and the exectution are to my opinion a little difficult to interpret. Two scoring systems, which are used for other purposes were used to quantify infiltrates in the lung on the first postop. day and this was used to develop a predictive score based on observations in less than 20 patients?
Reply 1: Thank you for your kind words and constructive feedback on our manuscript. We appreciate your interest and expertise in our topic. Below are our detailed responses to your comment.

We agree that the RALE and Brixia scoring systems were originally developed and validated for ARDS and COVID-19. However, we have learned that some studies have applied the RALE score to the evaluation of patients after lung, esophagus, and heart surgery. For example, Lizhen Xuan et al (https://doi.org/10.21037/jtd-23-822). used the RALE score to assess the severity of bilateral lung injury in patients after lung cancer surgery. This article was published in the JTD journal in October 2023. Xiaoliang Leng et al (https://doi.org/10.1053/j.semtcvs.2021.03.033). applied the RALE score to the postoperative evaluation of esophageal cancer patients and proposed a new method to identify acute lung injury (ALI). Karim Mostafa et al (https://doi.org/10.3390/jcm12186043). considered the RALE scoring system to be a practical tool that can objectively assess unilateral pulmonary edema (UPE) on chest X-ray images after mitral valve surgery (MVS). Therefore, we believe that these scoring systems are also applicable and useful for assessing postoperative pulmonary

complications in elderly patients after thoracic surgery. As for the Brixia score, there is no literature to date that indicates that the Brixia score is not suitable for assessing the severity of pneumonia in postoperative patients. Therefore, we only used the Brixia score for postoperative patients in an exploratory manner.

We acknowledge that our sample size was relatively small, as we only included 19 patients who experienced COT failure and required escalation to HFNC therapy. However, this reflects the actual proportion of COT failure in low-risk elderly patients, which was about 8% in our center. And we admit that a larger sample size would increase the statistical power and generalizability of our results, and we plan to conduct a prospective multicenter study in the future to validate our findings.

Changes in the text: We have modified our text as advised (see line 125-126).

Comment 2: It is well known that around 10% of our patients will develop postop. respiratory failure and we very well know that thoracotomy, surgical time etc. are risk factors for respiratory failure. And it is pretty sure that if the authors checked factors like preop lung function and exercise capacity, these factors would also make it to the score system. I am finding it difficult to accept the clinical use of an additional scoring system which is made up of already known risk factors and lung infiltrates.

Reply 2: We thank the reviewer for this comment. We agree that there are several well-known risk factors for postoperative respiratory failure, such as preoperative lung function, and exercise capacity. But none of these risk factors have a quantitative score. We just hope that our proposed scoring system can combine the known risk factors to alert the clinicians, and identify the patients who need to upgrade their oxygen therapy in advance. And in developing countries, premature adoption of HFNC not only places an economic burden on patients but also garners attention from medical insurance. We can identify the patients who need to upgrade their oxygen therapy without increasing their extra burden, just by reviewing the chest radiographs after surgery. Therefore, we think that our proposed scoring system has some clinical use.

Reviewer D

While your manuscript provides valuable insights into an important clinical topic, there are some areas that could be refined to further enhance the quality and impact of the work. Here are some respectful suggestions that could potentially improve the paper if you choose to implement them:

Title
- Consider writing in full the abbreviation COT (i.e., continuous oxygen therapy) since this is not a commonly used abbreviation and readers may not be familiar with the term.
Reply 1: Thank you for your comment. We agree that COT is not a widely used abbreviation and may cause confusion for some readers. Therefore, we have revised the title to spell out COT as continuous oxygen therapy.
Changes in the text: We have modfied our title as advised (line 2).

Keywords
- Consider including the phrase "chest radiograph" or similar words to increase the discoverability of your article.
Reply 2: Thank you for your valuable suggestion. We agree that adding the phrase "chest radiograph" or similar words to the keywords can improve the visibility and accessibility of our article.

Changes in the text: We have modfied our title as advised (line 73).

Introduction
- When citing previous studies, including brief mentions of their sample sizes and designs could provide helpful context about the strength of evidence being referenced.
- To make the objectives more apparent for readers, you could highlight and elaborate on the specific study aims in a separate paragraph.
Reply 3: Thank you for your valuable comments and suggestions. We have added the sample sizes and designs of the previous studies that we cited in the introduction, to provide more context and evidence for our research, and we have also highlighted and elaborated on our specific study aims in a separate paragraph at the end of the introduction, to make them more clear and explicit for the readers.
Changes in the text: We have modified our text as advised (see line 128-130, line 146-148).

Methods
- For the chest X-ray scoring process, elaborating on the training of the radiologists and experience of the senior radiologist would add credibility regarding quality control. What was the process for resolving any disagreements between the radiologists?
- Describing any statistical power calculations performed during study planning could demonstrate awareness about avoiding false negative findings.
Reply 4: Thank you for your valuable comments and suggestions. We have incorporated details regarding the training process and experience of each radiologist. Additionally, we have provided information on the procedure employed to resolve any disagreements among the radiologists. For the statistical power analysis, we conducted a thorough analysis at the early stages of study planning. However, in the clinical setting, the enrolled patients represent the entirety of the available population for this study.
Changes in the text: We have modified our text as advised (see line 196-201).

Results
- As positive and negative findings are both important, you could consider giving comparative emphasis.
- Incorporating visual plots like box plots for continuous variables comparisons could improve graphical representation for readers.
Reply 5: Thank you for your comments. We agree that positvie and negative findings are both important and we will incorporate box plots to better illustrate the distribution of the data. Such graphical representation will aid readers in a more intuitive understanding of the study results, providing a more comprehensive comparative

analysis.

Changes in the text: We have modifed our text as advised (see line 251-253).

Discussion

- The discussion interprets the results well, aligned with study aims. To augment the impact for clinicians, you could elaborate further on potential practice and policy implications.

Reply 6: We agree that this is an important aspect of our study and have added a paragraph in the discussion section to address the potential practice and policy implications of our findings.

Changes in the text: We have modifed our text as advised (see line 307-311).

- Commenting on the clinical applicability of the cut-off values identified for the scoring systems could better translate the predictive modelling for end users.

Reply 7: We thank the reviewer for this suggestion and have added a sentence in the discussion section to comment on the clinical applicability of the cut-off values identified for the scoring systems.

Changes in the text: We have modifed our text as advised (see line 359-364).

- Discussion strengths already highlight clinical relevance. Addressing limitations more substantially may further establish rigor and maturity of perspective.

Reply 8: We acknowledge that our study has some limitations and have expanded the discussion of the limitations in the discussion section.

Changes in the text: We have modifed our text as advised (see line 377-382).

- Describing any current barriers to adoption for the approach in real-world practice would demonstrate a thoughtful perspective about clinical integration.

Reply 9: Thank you for your valuable comment. We agree that it is important to discuss the potential barriers to the adoption of high-flow nasal cannula (HFNC) therapy in low-risk elderly patients after thoracic surgery. We have described the current barriers to adoption for the approach in real-world practice in the limitation section.

Changes in the text: We have modifed our text as advised (see line 377-382).

- Discussing any plans already underway to validate the model in larger datasets or other centres would showcase meaningful continuity of the research.

Reply 10: We agree that this is a relevant and important point and have added a sentence in the discussion section to discuss our plans.

Changes in the text: We have modifed our text as advised (see line 378-382).

Conclusions

- Including subtle nuance by using words like "may be predictive" rather than definitive statements could convey an appropriate degree of caution about generalizability.

Reply 11: Thank you for your valuable comments. We agree that our conclusions should be more cautious and nuanced, considering the limitations of our study. We have revised

our conclusions accordingly, using words like "may" and "suggest" to indicate the uncertainty and need for further validation.
Changes in the text: We have modfied our text as advised (see line 387).

- You may suggest more targeted next steps for research to provide continuity and advance the field.
Reply 12: Thank you for your comment. We have also added some suggestions for future research directions.
Changes in the text: We have modfied our text as advised (see line 392-393).

Incorporating some or all these additions can help bolster methodologic, analytic, and interpretive aspects for readers. By addressing subtle nuances, you can showcase scientific maturity to match the meaningful contributions of your work. This would further augment publication potential and impact for this manuscript.

Reviewer E

Thanks for submitting your paper to our attention. With this retrospective analysis you aimed to assess the efficacy of two chest X-ray scores in predicting continuous oxygen13 therapy treatment failure in patients over 70 years of age after thoracic surgery. The study is interesting and well written.
I have some minor comments:
1- the acronyms should be specified when comparing for the first time in the text (ROC, AUC, HFNC, etc).
Reply 1: Thank you for your comment. We agree that the acronyms should be explained when they are first used in the text.
Changes in the text: We have added the full names of the acronyms in parentheses after their first appearance in the text .

2 - I would suggest to avoid acronym in the title
Reply 2: Thank you for your suggestion. We agree that the title should be clear and concise without using acronyms.
Changes in the text: We have revised the title as follows: "The Chest X-ray Score Baseline in Predicting Failure of Continuous Oxygen Therapy in Low-Risk Elderly Patients After Thoracic Surgery".

3- Abstract: please rephrase the conclusions
Reply 3: Thank you for your comment. We have rephrased the conclusions to make them more specific and concise.
Changes in the text: We have changed the text as advised (see line 66-71).

4- line 106: exclusion instead of xclusion
Reply 4: Thank you for pointing out this typo. We have corrected it.
Changes in the text: We have changed "xclusion" to "exclusion" in line 163.

5 - line 106 and 107 pre-operative instead of Pre-operative

Reply 5: Thank you for pointing out this inconsistency. We have standardized the spelling of pre-operative.

Changes in the text: We have changed "Pre-operative" to "preoperative" in line 164.

6 - line 249: it's not clear for me what the approach "thoraco-abdominal laparoscopy" is.

Reply 6: Thank you for your comment. Thoraco-abdominal laparoscopy is a minimally invasive surgical technique that combines thoracoscopy and laparoscopy to access both the thoracic and abdominal cavities through small incisions. It is mainly used for esophageal cancer surgery, as it allows the resection of the tumor and the reconstruction of the digestive tract without opening the chest or abdomen.