Peer Review File

Article information: https://dx.doi.org/10.21037/jtd-22-1076

Reviewer A

The Authors performed a retrospective study in a sample of 387 hospitalized patients with COVID-19 who underwent chest CT and investigated whether quantitative CT lung data may help to predict respiratory outcomes including pneumonia, hypoxia and respiratory failure.

The design is not particularly new and the study suffers from several flaws.

Major

Comment 1. The Authors state that “Some patients underwent chest CT, and the need for chest CT was determined by each attending physician”.

The number of patients hospitalized with COVID-19 who did not undergo chest CT in the same period should be provided. Also, reasons to undergo CT should be better explained and accounted for. In fact, even in hospitalized subjects, presumably, in some instances CT was required to make diagnosis of COVID-19, in others CT was requested for clinical worsening in patients with already diagnosed COVID-19 pneumonia. This heterogeneity is confirmed in table 1 in which 10% of patients were qualified as “asymptomatic infection” and 4.9% as patients with “respiratory failure”. Clearly these situations are much different and the real generalizability of the results cannot escape a better description of this clinical sample.

Reply 1. We appreciate your great comments. Since this was a retrospective cohort study, study patients were heterogeneous in baseline characteristics including whether they underwent chest CT examination at the time of hospitalization. There was no required protocol for chest CT scans in cases of COVID-19 in our hospital and the need for chest CT was determined by each treating physician. The reasons for not performing the chest CT scan included confirmation of pneumonia with chest radiograph alone or difficulty of performing chest CT due to unstable vital signs. In contrast, the reasons for performing the chest CT scan included unclear diagnosis of pneumonia with chest radiographs alone, need for more precise evaluation of the extent of pneumonia, or suspicion of pulmonary embolism.

In our study, 147 patients underwent chest CT whereas 240 patients did not during the study period. In patients with pneumonia, hypoxia, and respiratory failure, 104 (53.3%), 56 (65.9%) and 15 (78.9%) underwent chest CT, respectively. As you pointed out, this heterogeneity may cause difficulty in interpreting the outcomes and limit generalizability of the study results. This limitation was inevitable due to the retrospective nature of our study; therefore, we believe that our approach with integrative prediction models should be tested in other patient groups. We have added this as our limitation in the Discussion section. We have also addressed characteristics of patients who underwent chest CT and who did not in the Result section. Please refer to the page 9, lines 190-192 and pages 14, lines 300-307 in our revised manuscript and Supplementary file.

Changes in the text: Among the study patients, 147 patients underwent chest CT whereas 240 patients did not at the time of diagnosis. The baseline characteristics between patients with and without chest CT scan are compared in Supplemental Table S1.

(Page 9, lines 190-192)

First, due to the retrospective nature of the study, Chest CT was not regularly performed in all study patients. The reasons for performing chest CT included unclear diagnosis of pneumonia with chest radiographs alone, need for more precise evaluation of the extent of pneumonia, or suspicion of pulmonary embolism. In the rest of the patients, chest CT was not performed because pneumonia was evident with chest radiograph alone or patients' unstable vital signs did not allow chest CT examinations. This heterogeneity may limit generalizability of the study results. Our approach with integrative prediction models should be tested in further studies.

(Page 14, lines 300-307)

Comment 2. In the description of quantitative chest analyses the Authors state that the high attenuation areas in the -600 to -250 HU range were considered to represent the extent of pneumonia. However, by including only areas between -600 and -250HU you seem to deliberately excluded areas of lung consolidation from the analysis. Lung consolidation is largely present in COVID-19 and it is correlated with mortality and ICU admission (doi: 10.1007/s11547-020-01305-9; doi: 10.1007/s00330-020-07033-y;

doi: 10.1038/s41598-021-95114-3) Why were consolidations not included in the analysis? If lung consolidations are excluded from the analysis, this is a noteworthy limitation of the study.

Reply 2. We appreciate your great comment. HAA is the lung volume containing CT attenuation values higher than those of normal lung parenchyma. The extent of HAA is very low in normal lungs and represents the densities of physiological structures (e.g., bronchial walls and vessels). Although the term was first conceived in the interstitial lung disease, HAA also corresponds to superimposed disease-related alterations such ground glass opacities and consolidations, reflecting the extent of the disease (doi: 10.3390/diagnostics11050738). However, we agree with your opinion that the threshold used for HAA would not capture areas with dense consolidation. Considering the threshold for consolidation used in a previous study, we measured the extents of areas with -100 and 0 Hounsfield units and built prediction models with them. With the software we used, automated quantification of areas with Hounsfield units greater than 0 was not available. We have added the results of new analyses in the revised manuscript. Please refer to the following contents and revised manuscript.

Changes in the text: Whole-lung images were extracted from the chest wall, mediastinum, and large airways, and attenuation coefficients of pixels were measured sequentially for indexes including the quantified percentage of low-attenuation area (LAA) less than -950 Hounsfield units (HU), high-attenuation area (HAA) between -600 and -250 HU, and consolidation between -100 and 0 HU using a multilayer convolutional neural network.

(Page 8, lines 153-158)

Consolidation (%) was significantly higher in patients with pneumonia and hypoxia.

(Pages 10-11, lines 222-223)

Neutrophil and lymphocyte percentages and levels of AST, CRP, HAA (%), and consolidation (%) were associated with all three respiratory outcomes.

(Page 11, lines 231-233)

To predict hypoxia, the presence of hypertension and levels of LDH, CRP, HAA (%), and consolidation (%) were chosen.

(Page 11, lines 235-237)

The top 10 predictors for pneumonia were ferritin, CRP, fibrinogen, platelet count, neutrophil percentage, HAA (%), LDH, age, vaccination status, and WBC; predictors for hypoxia were LDH, CRP, neutrophil percentage, fibrinogen, procalcitonin, ferritin, HAA (%), LAA (%), lymphocyte percentage, and AST; and predictors for respiratory failure were HAA (%), CRP, LDH, AST, procalcitonin, Ct value of RdRp gene, ferritin, presence of chronic kidney disease, neutrophil percentage, and body mass index. A random forest model was developed for each respiratory outcome.
(Pages 11-12, lines 247-254)

Comment 3. The phase of the COVID-19 epidemic (wave, agent etc.) in the Sept to Dec 2021 time frame in the country where the study was performed (I presume South Korea) should be provided.
Reply 3. Thank you for an important comment. We have added the information in the Methods section as advised. Please refer to the page 6, lines 106-110, in the revised manuscript.
Changes in the text: Although genotyping of SARS-CoV-2 was not performed in our study patients, the Delta variant may have been the predominant type among the study patients because the detection rate of the Delta variant was greater than 50% of the local cases in our country by the end of July 2021. The Omicron variant had not yet become the dominant variant until January 2022.
(Page 6, lines 117-121)

Comment 4. Validation of the models should be tested in an independent sample to avoid a "peeking effect".
Reply 4. We agree with your concern that the model validation should be performed in an external dataset. We attempted to validate our data using 10-fold cross-validations. However, it has been recognized that testing a model using a dataset that is already used in a pre-processing stage such as feature selection can lead to overestimation of generalizability of the classifiers, also known as a "peeking effect" (doi: 10.3174/ajnr.A3685). Unfortunately, we were unable to test our models in external patient samples because the current patient cohort was the complete list of patients

obtained for this study approved by the institutional review board. Further, due to the small number of patients in our cohort, it was limited to divide patients into a training set and testing set. We believe that further large-scale, prospective studies need to be conducted to confirm the findings of our study. According to your opinion, we have added this as a limitation of our study in the Discussion section. Please refer to the page 15, lines 323-326 in the revised manuscript.

Changes in the text: Lastly, our models were not validated externally. Although we performed cross-validation, overestimation of the generalizability might have occurred. Further studies need to validate our models using new data from different settings.

(Page 15, lines 323-326)

Minor

Comment 5. Abstract: The results section is not clear. In the M&Ms you write about a model including demographic, laboratory and CT findings, while in the results you speak predominantly about high attenuation areas, without having defined them accurately in the previous section. Demographics and laboratory findings are also not mentioned.

Reply 5. We appreciate your valuable comments. According to your opinion, we have added the missing information in the Methods and Results section in our abstract. Since we have added the contents about consolidation, we have also included the definition of consolidation in the abstract. Please refer to the revised manuscript page 3.

Changes in the text:

Methods: High-attenuation area (HAA) (%) and consolidation (%) were defined as quantified percentages of the area with Hounsfield units between -600 and -250 and between -100 and 0, respectively.

(Page 3, lines 46-48)

Results: A total of 195 (50.4%), 85 (22.0%), and 19 (4.9%) patients developed pneumonia, hypoxia, and respiratory failure, respectively. The mean patient age was 57.8 years, and 194 (50.1%) were female. In the multivariable analysis, vaccination status and levels of lactate dehydrogenase, C-reactive protein (CRP), and fibrinogen were independent predictors of pneumonia. The presence of hypertension, levels of lactate dehydrogenase and CRP, and HAA (%) were selected as independent variables to predict hypoxia. For respiratory failure, the presence of diabetes, levels of aspartate aminotransferase, and CRP, and HAA (%) were selected.

(Page 3, line 54-61)

Comment 6. Methods and results: Were HAA expressed as a percentage or as an absolute value?

Reply 6. Thank you for pointing out an ambiguous part. The HAA was expressed as a percentage of the whole lung volume (%). We have clarified all the parts. Please see the revised manuscript.

Comment 7. The Authors utilized a chest low dose protocol (120 100 kVP and current of 24 mA). While this has no impact on the definition of emphysema when measured with LAA 950, I am not sure that such a technical choice is not without consequences for assessment of HHA. Please check.

Reply 7. Thank you for pointing out an important point. We have checked the CT protocol and the 50 mA is used for usual chest CT scans. Therefore, we have corrected the information accordingly. Still, 50 mA is considered a low-dose protocol current. Given that HAA was first mentioned in the context of interstitial lung disease (ILD), high resolution CT is the preferred method to measure HAA. In fact, high resolution CT is the gold standard imaging modality for the diagnosis of ILD. Many recent studies have focused on establishing the optimal low dose CT protocol and tube current as low as 20 mA was found appropriate for diagnosing ILD (doi: 10.1016/j.pulmoe.2020.06.004). Our CT images were reconstructed with 1.0 mm thickness which is an important determinant in image quality. In addition, the CT analysis software was customized for our CT protocol. Therefore, we believe HAA quantification with our protocol was valid.

Changes in the text:

Chest CT images were obtained using standardized CT screening protocols at a tube voltage of 120 kVP and current of 50 mA, which were applied in the high-pitch spiral mode (Aquilion One, Toshiba).

(Page 8, lines 148-150)

Reviewer B

Comment 1. The major flaw of this manuscript is lack of external validation after the developing the model. External validation is a necessary for this research.

Reply 1. We agree with your concern that the model validation should be performed in an external dataset. We attempted to validate our data using 10-fold cross-validations. However, it has been recognized that testing a model using a dataset that is already used in a pre-processing stage such as feature selection can lead to overestimation of generalizability of the classifiers, also known as a “peeking effect” (doi: 10.3174/ajnr.A3685). Unfortunately, we were unable to test our models in external patient samples because the current patient cohort was the complete list of patients obtained for this study approved by the institutional review board. Further, due to the small number of patients in our cohort, it was limited to divide patients into a training set and testing set. We believe that further large-scale, prospective studies need to be conducted to confirm the findings of our study. According to your opinion, we have added this as a limitation of our study in the Discussion section. Please refer to the page 15, lines 323-326 in the revised manuscript.

Changes in the text: Lastly, our models were not validated externally. Although we performed cross-validation, overestimation of the generalizability might have occurred. Further studies need to validate our models using new data from different settings.

(Page 15, lines 323-326)

Comment 2. Besides, there is no unified standard on when to perform chest CT for patients with COVID-19. This may introduce bias to the data set.

Reply 2. We appreciate your great comment. Since this was a retrospective cohort study, study patients were heterogeneous in baseline characteristics including whether they underwent chest CT examination at the time of hospitalization. There was no required protocol for chest CT scans in cases of COVID-19 in our hospital and the need for chest CT was determined by each treating physician. The reasons for not performing the chest CT scan included confirmation of pneumonia with chest radiograph alone or difficulty of performing chest CT due to unstable vital signs. In contrast, the reasons for performing the chest CT scan included unclear diagnosis of pneumonia with chest radiographs alone, need for more precise evaluation of the extent of pneumonia, or suspicion of pulmonary embolism.

In our study, 147 patients underwent chest CT whereas 240 patients did not during the study period. In patients with pneumonia, hypoxia, and respiratory failure, 104 (53.3%), 56 (65.9%) and 15 (78.9%) underwent chest CT, respectively. As you pointed out, this heterogeneity may cause difficulty in interpreting the outcomes and limit generalizability of the study results. This limitation was inevitable due to the retrospective nature of our study; therefore, we believe that our approach with integrative prediction models should be tested in other patient groups. We have added this as our limitation in the Discussion section. We have also addressed characteristics of patients who underwent chest CT and who did not in the Result section. Please refer to the page 9, lines 190-192 and pages 14, lines 300-307 in our revised manuscript and Supplementary file.

Changes in the text: Among the study patients, 147 patients underwent chest CT whereas 240 patients did not at the time of diagnosis. The baseline characteristics between patients with and without chest CT scan are compared in Supplemental Table S1.

(Page 9, lines 190-192)

First, due to the retrospective nature of the study, Chest CT was not regularly performed in all study patients. The reasons for performing chest CT included unclear diagnosis of pneumonia with chest radiographs alone, need for more precise evaluation of the extent of pneumonia, or suspicion of pulmonary embolism. In the rest of the patients, chest CT was not performed because pneumonia was evident with chest radiograph alone or patients' unstable vital signs did not allow chest CT examinations. This heterogeneity may limit generalizability of the study results. Our approach with integrative prediction models should be tested in further studies.

(Pages 14, lines 300-307)

Reviewer C

Comment 1. What is your purpose for prediction of respiratory outcomes of COVID-19? This point can be mentioned in the section of INTRODUCTION.

Reply 1. Thank you for your valuable comment. Most patients with COVID-19 infection experience a mild disease course while some require hospitalization or even progress to severe respiratory failure. Timely detection of high-risk patients is imperative for delivering proper management while optimizing the use of limited resources. With the help of artificial intelligence, early detection of these high-risk patients will help in establishing a patient management plan. Patients' outcomes can be predicted in a simple, easy, and fast way by utilizing the radiologic parameters obtained from an automated CT quantification program. We have added the following sentence in the Introduction section to clarify our purpose. Please see page 6, lines 108-109.

Changes in the text: This integrative model may enable simple and fast identification of high-risk patients at an early stage of the disease.

(Page 6, lines 108-109)

Comment 2. Why did you use two machine learning models (logistic regression) to show the prediction results?

Reply 2. Thank you for your valuable comment. As you pointed out, we presented two analyses: logistic regression and random forest modelling. In the logistic regression analysis, we created prediction models with significant variables obtained from unadjusted and adjusted analyses and corresponding ROC curves were constructed. While logistic regression model is a more traditional analysis method, the random forest is one of commonly used machine-learning algorithms in which it is not possible to know specifically how the result was obtained, therefore also referred as a black-box machine learning model. Random forest was performed to find variables significantly associated with the respiratory outcomes using the feature selection which is a completely different method from logistic regression. By diversifying the analysis methods, we tried to improve the quality and reliability of the analysis results.

Comment 3. One limitation must be mentioned. There is no external validation dataset.

Reply 3. We agree with your concern that the model validation should be performed in an external dataset. We attempted to validate our data using 10-fold cross-validations. However, it has been recognized that testing a model using a dataset that is already used in a pre-processing stage such as feature selection can lead to overestimation of generalizability of the classifiers, also known as a “peeking effect” (doi: 10.3174/ajnr.A3685). Unfortunately, we were unable to test our models in external patient samples because the current patient cohort was the complete list of patients obtained for this study approved by the institutional review board. Further, due to the small number of patients in our cohort, it was limited to divide patients into a training set and testing set. We believe that further large-scale, prospective studies need to be conducted to confirm the findings of our study. According to your opinion, we have added this as a limitation of our study in the Discussion section. Please refer to the page 15, lines 323-326 in the revised manuscript.

Changes in the text: Lastly, our models were not validated externally. Although we performed cross-validation, overestimation of the generalizability might have occurred. Further studies need to validate our models using new data from different settings.

(Page 15, lines 323-326)