Peer Review File

Article information: https://dx.doi.org/10.21037/atm-21-5571

Reviewer A

The authors present a multimodal approach for COVID-19 diagnosis which combines conventional clinical factors and CT imaging and takes advantage of machine learning models.

The strengths of the manuscript:

1. The manuscript is well-written.

2. The cohort is multi-centric and inclusive enough with over 700 subjects.

3. The train/test split with 7:3 ratio is fine.

Response:
We thank the reviewer for the praise.

The weaknesses of the manuscript:
1. The details of the distribution of the patients in the 11 study sites is missing. E. g,, the number of patients in each center.
2. The scanning protocols and resolutions of the CT scanners at each center is missing.

Response:
Thank you for pointing this out. We have added the details in Table 2 and revised the results as follows (page 15).

Patient characteristics and CT scanning protocols are shown in Tables 1 and 2.

Table 2. Patient Characteristics and scanning protocol in each hospital

Characteristic	H01 N = 94[1]	H02 N = 158[1]	H03 N = 19[1]	H04 N = 70[1]	H05 N = 71[1]	H06 N = 32[1]	H07 N = 21[1]	H08 N = 68[1]		H09 N = 110[1]	H10 N = 34[1]	H11 N = 43[1]
Age	57 (44, 76)	60 (44, 73)	63 (53, 76)	60 (39, 72)	59 (44, 76)	60 (46, 68)	65 (41, 76)	63 (42, 73)		59 (42, 73)	78 (58, 86)	68 (46, 80)
COVID-19 PCR												
Positive	70 (74%)	52 (33%)	18 (95%)	35 (50%)	26 (37%)	21 (66%)	12 (57%)	18 (26%)		37 (34%)	11 (32%)	21 (49%)
System	Aquilion PRIME	Optima CT660	Aquilion PRIME	Optima CT660	Optima CT660	Aquilion PRIME	Aquilion CX Edition	Aquilion ONE	Aquilion CXL	Aquilion PRIME	Aquilion CXL	Aquilion CX Edition
Vendor	Canon Medical Systems	GE	Canon Medical Systems	GE	GE	Canon Medical Systems	Canon Medical Systems	Canon Medical Systems	Canon Medical Systems	Canon Medical Systems	Canon Medical Systems	Canon Medical Systems

Characteristic	H01 N = 94[1]	H02 N = 158[1]	H03 N = 19[1]	H04 N = 70[1]	H05 N = 71[1]	H06 N = 32[1]	H07 N = 21[1]	H08 N = 68[1]		H09 N = 110[1]	H10 N = 34[1]	H11 N = 43[1]
Tube voltage (kVp)	120	120	120	120	120	120	120	120	120	120	120	120
Automatic tube current Modulation (mAs)	Auto	100-510	150-250	80-500	80-500	150-250	403-500	100-400	100-400	50-250	100-400	100-400
Pitch												
Standard	111	55	65	55	55	65		65	53	65	53	
Factor	0.813	0.984	0.813	0.984	0.984	0.813	1.172	0.813	0.828	0.813	0.828	1

Characteristic	H01 N = 94[1]	H02 N = 158[1]	H03 N = 19[1]	H04 N = 70[1]	H05 N = 71[1]	H06 N = 32[1]	H07 N = 21[1]	H08 N = 68[1]		H09 N = 110[1]	H10 N = 34[1]	H11 N = 43[1]
Matrix	512x512	512x512	512x512	512x512	512x512	512x512	512x512	512x512	512x512	512x512	512x512	512x512
Slice thickness (cm)	0.5	0.625	0.5	0.625	0.625	0.5	0.5	0.5	0.5	0.5	1.0	5.0
Field of view (mm)	320	340	320			330	350	320	320	320-400	320	320
Reconstruction interval (mm)	5	0.625	2	1.25	1.25	3	5	5	5	5	5	5

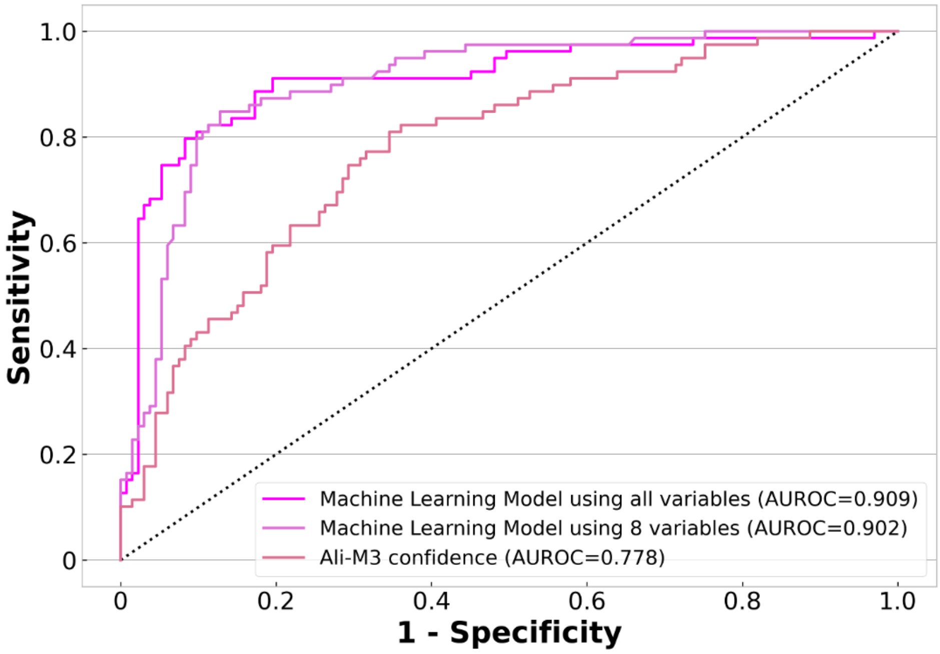
[1] n (%); Median (IQR)

3. It is unclear, how the data from 11 centers are aggregated. If the data are aggregated without proper normalization, then the train/test methodology will be invalidated. If the CT acquisition methodologies differ between centers, the train/test cohorts cannot be aggregated by simply appending them together. Instead:
3.a. for each center, the exact acquisition methodology should be described in methods section.
3.b. if already applied, the details of normalization of the dataset should be included. If not, proper normalization technique should be applied, considering random effects exposed by the varying centers.

Response:
We understand the reviewer's concern regarding normalization. The CT acquisition methodologies differ between centers. However, we did not apply any normalization methodology because, at the development stage of Ali-M3, the developers did not normalize for each CT machine. Ali-M3 accepts 512*512-pixel DICOM images. We simply applied Ali-M3 to our dataset. This decision would help us estimate the accuracy of our model when applied to hospitals not included in this study.

4.A Receiver Operating Characteristic (ROC) curve plots true positive rate values (i. e., sensitivity) against true negative rate values (1 - specificity) at different thresholds. However, the X axis of the ROC curve presented in figure 1 shows specificity. I assume, it has been a typo and it should be 1 - specificity. Otherwise, the diagram shows values different than expected. In either case, this should be clarified and revised accordingly.

Response:
We showed the X axis from 1 to 0. Hence, the X axis correctly showed the 1 – specificity. We have revised the figure to avoid misunderstanding.



Reviewer B

The authors developed and validated a machine learning diagnostic model for novel coronavirus (COVID-19) disease, integrating artificial-intelligence-based computed tomography (CT) imaging and clinical features. This paper deals with a hot area of investigation at the moment. However, the study lacks a clear comparison between the submitted paper and the more relevant literature contributions, which should highlight the main advantages of the current submission. Can be added to the reference section be more appropriate.
DOI: 10.26355/eurrev_202009_22875
DOI: 10.26355/eurrev_202011_23640
DOI: 10.26355/eurrev_202008_22510

Response:
We thank the reviewer for the constructive comments. We have read through the list of recommended references and added only the most relevant paper into Introduction (page 10).

> Consequently, there are no diagnostic models using chest computed tomography (CT) with potential clinical use (9,10).

Reviewer C

This study describes the validation of the machine learning diagnostic model for COVID-19. The authors retrospectively analysed the cohort dataset collected in 11 tertiary care facilities. The authors concluded that the combination of machine learning and CT evaluation with blood could be used for the rapid diagnosis of COVID-19.

OVERALL IMPRESSION
This study addresses some interesting points by showing that the diagnostic accuracy for COVID-19 is improved by adding blood test results to the existing AI-based CT image analysis results.
My main concern is that the WHO has recommended limited use of imaging tests for the diagnosis of COVID-19, and chest CT also has limited diagnostic value for COVID-19, particularly in mild, or asymptomatic patients.

Response:
We thank the reviewer for the constructive comments. We intend to use this model in emergency departments for patients with severe symptoms. Of course, asymptomatic or mild patients do need not to draw a blood sample or to take a CT. To clarify this point, we have revised the discussion. Please see the response page 8 of this document.

SECTION-BY-SECTION

Methodology
- line 197 Participants – Please provide additional descriptions of the inclusion criteria. Were the study participants the emergency department visit patients, outpatients, or admitted patient? What was the suspected symptom? Please describe the inclusion criteria of the study participants. It would be easier to understand if the flow chart of the study participants was presented according to the inclusion and exclusion criteria.

Response:
We understand the reviewer's concern. Due to the retrospective nature, we were not able to define rigorous inclusion criteria. Patients who underwent RT-PCR and CT imaging in the chaotic conditions of the 1st wave of the COVID-19 pandemic were the participants of this study. These criteria included all emergency department visit patients, outpatients, and admitted

patients. The reviewer can access the flowchart in the previous publication.
https://journals.plos.org/plosone/article/figure/image?size=inline&id=10.1371/journal.pone.0258760.g001
We decided not to add this figure in our manuscript to avoid duplicate publication. We have added the limitation as follows (page 17–18):

> Third, because of the retrospective nature, we could not define rigorous inclusion criteria, such as symptoms or settings.

- line 207 Chest CT and Artificial intelligence – Please briefly and clearly describe which deep-learning model the authors used. It is not described in the references added by the author, and the appendix of the reference is not provided.

Response:
We understand the reviewer's concern. We published the manuscript in *PlosOne* this November. The reviewer can access the details with the following link:
https://doi.org/10.1371/journal.pone.0258760.s007
We have updated the reference from medRxiv to *PlosOne*.

- Page 269, Model validation – For model development, the ratio of the training set and the test set was 7:3. And the bootstrapped resampling (1000 samples) method was used for the validation set. If the data that development and validation data were used the same participant data set, is there any possibility that this result would be overestimated?

Response:
We understand the reviewer's concern. We chronologically split the dataset. Hence, the development and validation dataset were not duplicated. Of course, we believe that further updates are necessary to reflect the recent situation.

Discussion
The authors state that the authors diagnostic model automatically interpret clinical data in conjunction with CT scans. Also, the authors reported that several problems such as separate collection of cases and controls, lack of external validation, and insufficient reporting have been overcome in this study with rigorous methodology, with our model achieving good discrimination and calibration performance. And, the authors concluded that the A-blood model would allow for quicker diagnoses, and even if the RT-PCR test existed in the facility, the A-blood model would be a better option.

If I understood the methodology correctly, no external validation was performed in this study.

Response:
We understand the reviewer's concern and thus used the term "external validation" as the meaning of Type 2b.
https://www.acpjournals.org/na101/home/literatum/publisher/acp/journals/content/aim/2015/aim.2015.162.issue-1/m14-0698/20211006/images/medium/2ff4_figure_1_types_of_prediction_model_studies_covered_by_the_tripod_statement.jpg

In the TRIPOD statement, the external validation was defined as follows:

> External validation may use participant data collected by the same investigators, typically using the same predictor and outcome definitions and measurements, but sampled from a later period (temporal or narrow validation).

If the reviewer used "external validation" as only the meaning of Type 3, we agree with the reviewer's comment. We have added an explanation in the method section to clarify this point as follows (page 15):

2.7 Model External Validation

We used the temporal validation method for external validation.

No objective time variable data were provided that A-Blood model enable a faster diagnosis for COVID-19.

Response:

We agree with the reviewer's concern. We have weakened the expression and revised the discussion as follows (page 17):

> The A-blood model may allow for quicker diagnoses at emergency departments. Even if the RT-PCR test existed in the facility, the A-blood model might be a better option because of its lower turnaround time, which requires only a general blood test and CT results. In the majority of Japanese emergency hospitals, including the 11 hospitals in the dataset, the time to obtain CT imaging for stroke patients is less than 20 min (22). Even during the COVID-19 pandemic, no substantial increase was observed in the time to obtain CT (23).
>
> 22. Nationwide questionnaire survey on neuroimaging strategy for acute ischemic stroke in Japan. Japanese J Stroke [Internet]. 2020;42(6):502–8. Available from: https://www.jstage.jst.go.jp/article/jstroke/42/6/42_10781/_article/-char/ja/
> 23. Koge J, Shiozawa M, Toyoda K. Acute Stroke Care in the With-COVID-19 Era: Experience at a Comprehensive Stroke Center in Japan. Front Neurol [Internet]. 2021 Jan 18;11.
> Available from: https://www.frontiersin.org/articles/10.3389/fneur.2020.611504/full

And I would like to also ask the authors opinions whether the findings of this study can be generalized to other countries with different patient severities, medical resources, and environments.

Response:

We understand the reviewer's concern. First, Ali-M3 can only be used in mainland China and Japan. In another study currently under submission, poor CT performance tended to reduce the COVID-19 diagnostic accuracy of AI (data not shown). In both countries, the CT performance for general uses is similar. We suppose a certain degree of generalizability may exist. Of course, further "external validation" should be performed in other hospitals.