Title

Accuracy of deep learning based computed tomography diagnostic system of COVID-19: a consecutive sampling external validation cohort study

Authors

Tatsuyoshi Ikenoue, MD MPH PhD[1*], Yuki Kataoka, MD MPH DrPH[2,3*], Yoshinori Matsuoka, MD DrPH[4], Junichi Matsumoto, MD PhD[5], Junji Kumasawa, MD DrPH[6], Kentaro Tochitatni, MD[7], Hiraku Funakoshi, MD MPH PhD[8], Tomohiro Hosoda, MD PhD[9], Aiko Kugimiya, MD[10], Michinori Shirano, MD PhD[11], Fumiko Hamabe, MD[12], Sachiyo Iwata, MD PhD[13], Shingo Fukuma, MD PhD[1] and Japan COVID-19 AI team

*Contributed equally to this work with: Tatsuyoshi Ikenoue, Yuki Kataoka

Affiliations

[1]Human Health Sciences, Kyoto University Graduate School of Medicine, 53 Kawaharacho, Shogoin, Sakyo-ku Kyoto 606-8507, Japan

[2]Hospital Care Research Unit, Hyogo Prefectural Amagasaki General Medical Center, Higashi-Naniwa-Cho 2-17-77, Amagasaki, Hyogo 660-8550, Japan

[3]Department of Respiratory Medicine, Hyogo Prefectural Amagasaki General Medical Center, Higashi-Naniwa-Cho 2-17-77, Amagasaki, Hyogo 660-8550, Japan

[4]Department of Emergency Medicine, Kobe City Medical Center General Hospital, Minatojima-Minamimachi 2-1-1, Chuo-ku, Kobe-city, Hyogo 650-0047, Japan

[5]Department of Emergency and Critical Care Medicine, St. Marianna University School of Medicine, Sugao 2-16-1, Miyamae, Kawasaki, Kanagawa, 216-8511, Japan

[6]Department of Critical Care Medicine, Sakai City Medical Center, Nishi-Ku, Ebaraji-Cho 1-1-1,

Sakai, Osaka 593-8304, Japan

[7]Department of Infectious Diseases, Kyoto City Hospital, Mibu Higashi Takadacho 1-2, Nakagyo-ku, Kyoto-city, Kyoto 604-8845, Japan

[8]Department of Emergency and Critical Care Medicine, Tokyobay Urayasu Ichikawa Medical Center, Todaijima Urayasu Chiba 3-4-32, Japan

[9]Department of Infectious Disease, Kawasaki Municipal Kawasaki Hospital, Shinkawadori 12-1, Kawasaki-ku, Kawasaki, Kanagawa, 210-0013 Japan.

[10]Department of Emergency and Critical Care Medicine, Yamanashi Prefectural Central Hospital, Fujimi-Cho 1-1-1, Kofu, Yamanashi 400-8506, Japan

[11]Department of Infectious Diseases, Osaka City General Hospital, Miyakojima-Hondori 2-13-22, Miyakojima-ku, Osaka, 534-0021, Japan

[12]Department of Radiology, National Defense Medical College Hospital, Namiki 3-2, Tokorozawa, Saitama 359-0583, Japan

[13]Division of Cardiovascular Medicine, Hyogo Prefectural Kakogawa Medical Center, Kanno 203, Kanno-cho, Kakogawa 6758555, Japan

Group authors

1. Shingo Hamaguchi, MD, PhD

   Department of Emergency and Critical Care Medicine, St. Marianna University School of Medicine, Kanagawa, Japan

2. Takafumi Haraguchi, MD

   Department of Radiology, St. Marianna University School of Medicine, Kanagawa, Japan

3. Shungo Yamamoto, MD, DrPH

   Department of Infectious Diseases, Kyoto City Hospital, Kyoto, Japan

4. Hiromitsu Sumikawa, MD, PhD

   Department of Diagnostic Radiology, Sakai City Medical Center, Osaka, Japan

5. Koji Nishida, MD

Department of Respiratory Medicine, Sakai City Medical Center, Osaka, Japan

6. Haruka Nishida, MD

Department of Emergency Medicine, Kobe City Medical Center General Hospital, Hyogo, Japan

7. Koichi Ariyoshi, MD, PhD

Department of Emergency Medicine, Kobe City Medical Center General Hospital, Hyogo, Japan

8. Hiroshi Shinmoto, MD, PhD

Department of Radiology, National Defense Medical College Hospital, Saitama, Japan

9. Hiroaki Sugiura, MD, PhD

Department of Radiology, National Defense Medical College Hospital, Saitama, Japan

10. Hidenori Nakagawa, MD

Department of Infectious Diseases, Osaka City General Hospital, Osaka, Japan

11. Tomohiro Asaoka, MD

Department of Infectious Diseases, Osaka City General Hospital, Osaka, Japan

12. Naofumi Yoshida, MD, PhD

Division of Cardiovascular Medicine, Department of Internal Medicine, Kobe University Graduate School of Medicine, Kobe, Japan

13. Rentaro Oda, MD

Department of Infectious disease, Tokyobay Urayasu Ichikawa Medical Center, Chiba, Japan

14. Takashi Koyama, MD

Department of Infectious diseases, Hyogo Prefectural Amagasaki General Medical Center, Hyogo, Japan

15. Yui Iwai, MD

Department of Infectious diseases, Hyogo Prefectural Amagasaki General Medical Center, Hyogo, Japan

16. Yoshihiro Miyashita, MD

Department of Respiratory Medicine, Yamanashi Prefectural Central Hospital, Yamanashi, Japan

Corresponding author:

Tatsuyoshi Ikenoue

Human Health Sciences, Kyoto University Graduate School of Medicine, 53 Kawaharacho, Shogoin, Sakyo-ku Kyoto 606-8507, Japan

Telephone number: +81 -75-751-3925

E-mail address: ikenoue.tatsuyoshi.4e@kyoto-u.ac.jp

Funding information: None

Manuscript Type: Original Article

Word Count for Text: 2,696

Title

Accuracy of deep learning based computed tomography diagnostic system of COVID-19: a consecutive sampling external validation cohort study

Abstract:

Objectives: Ali-M3, an artificial intelligence, analyses chest computed tomography (CT) and detects the likelihood of coronavirus disease (COVID-19) in the range of 0 to 1. It demonstrates excellent performance for the detection of COVID-19 patients with a sensitivity and specificity of 98.5 and 99.2%, respectively. However, Ali-M3 has not been externally validated. Our purpose is to evaluate the external validity of Ali-M3 using Japanese sequential sampling data.

Methods: In this retrospective cohort study, COVID-19 infection probabilities were calculated using Ali-M3 in 617 symptomatic patients who underwent reverse transcription-polymerase chain reaction (RT-PCR) tests and chest CT for COVID-19 diagnosis at 11 Japanese tertiary care facilities, between January 1 and April 15, 2020.

Results: Of 617 patients, 289 patients (46.8%) were RT-PCR-positive. The area under the curve (AUC) of Ali-M3 for predicting a COVID-19 diagnosis was 0.797 (95% confidence intervals [CI]: 0.762–0.833) and goodness-of-fit was P = 0.156. With a cut-off of probability of COVID-19 by Ali-M3 diagnosis set at 0.5, the sensitivity and specificity were 80.6% and 68.3%, respectively, while a cut-off of 0.2 yielded a sensitivity and specificity of 89.2% and 43.2%, respectively. Among 223 patients who required oxygen support, the AUC was 0.825 and sensitivity at a cut-off of 0.5 and 0.2 were 88.7% and 97.9%, respectively. Although the sensitivity was lower when the days from symptom onset were few, sensitivity increased for both cut-off values after 5 days.

Conclusions: Ali-M3 was evaluated by external validation and shown to be useful to exclude a diagnosis of COVID-19.

Key words:

COVID-19; artificial intelligence; chest CT imaging; reverse transcription polymerase chain reaction; external validation study

Key Points:

1. The area under the curve (AUC) of Ali-M3, which is an AI system for diagnosis of COVID-19 based on chest CT images, was 0.797 and goodness-of-fit was P = 0.156.
2. With a cut-off of probability of COVID-19 by Ali-M3 diagnosis set at 0.5, the sensitivity and specificity were 80.6% and 68.3%, respectively, while a cut-off of 0.2 yielded 89.2% and 43.2%.
3. Although low sensitivity was observed in less number of days from symptoms onset, after 5 days high increasing sensitivity was observed. In patients requiring oxygen support, the AUC was higher that is 0.825.

Abbreviations:

AI = artificial intelligence

COVID-19 = coronavirus disease 2019

CT = computed tomography

Introduction

A proper triage system is necessary during this coronavirus disease (COVID-19) pandemic era,[1, 2] as improper triage systems may disadvantage patients and lead to wastage of personal protective equipment (PPE) and hospital infections through admission of infected patients to facilities, causing collapse of the medical system. Although reverse transcription-polymerase chain reaction (RT-PCR) tests have been developed, the delay in waiting for RT-PCR results can hamper proper triage.

Computed tomography (CT) is a fast and useful diagnostic tool. Some studies have reported the characteristic findings on chest CT images of COVID-19 patients.[3-8] Use of chest CT images by radiologists has shown high diagnostic performance for COVID-19. However, even radiologists' interpretations vary largely, because of the influence of their habituation in the interpretation of COVID-19 CT images.[9] Therefore, using CT as a diagnostic tool in general clinical practice is difficult in the current situation.

Diagnostic support systems using artificial intelligence (AI) have the potential to replace many of the routine detection, characterisation, and quantification tasks currently performed by radiologists using cognitive ability.[10] AI can prevent the variability of diagnosis from inter- and intra-reader variability. In China, where COVID-19 infection originated, many AI systems were developed for establishing a diagnosis of COVID-19 based on chest CT images.[11-15] One such system, Ali-M3, can detect the likelihood of COVID-19 in the range of 0 to 1, and has excellent accuracy for the detection of COVID-19 with an accuracy, sensitivity, and specificity of 99.0, 98.5, and 99.2%, respectively. Although Ali-M3 has excellent accuracy, it was developed in a virtual population, which consisted of 3,067 examinations for COVID-19; 1,996 for community-acquired pneumonia; and 1,975 for non-pneumonia, which was different from the general population and its accuracy could be overestimated.[16]

To use Ali-M3 to diagnose exclusion of COVID-19, its external validity must be evaluated based on the

distribution of diseases in a real-world setting. We here conducted a retrospective cohort study to evaluate the external validity of Ali-M3 using the Japanese sequential sampling data of patients who underwent RT-PCR tests and chest CT for diagnosis of COVID-19.

Materials and Methods

Study design

This retrospective cohort study consisted of 11 Japanese tertiary care facilities that provided treatment for COVID-19 in each area. We partially followed the guidelines of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis Statement to plan and report this study (Supplemental Table 1).[17] The institutional review board of each facility approved the study and the need to obtain written informed consent was waived.

Participants

We included patients who underwent both RT-PCR examinations and chest CT for the diagnosis of COVID-19. The potentially eligible participants were identified on the advice of physicians that both RT-PCR test and chest CT be obtained when the patients presented with symptoms or were suspected of having COVID-19. The detailed information of the inclusion criteria is shown in Supplemental Table 2. We selected patients by using consecutive sampling methods between January 1 and April 15, 2020. The RT-PCR results were extracted from the patients' medical records at each facility. Patients were excluded when the time-interval between chest CT and the first RT-PCR assay was longer than 7 days.

All available data on the database were used to maximize the power and generalizability of the results.

Chest CT protocols

All images were obtained on one of five types of CT systems, with the patient in the supine position. The details of scanning parameters and systems are shown in Supplemental Table 3.

Image analysis

We used a three-dimensional deep learning framework for the detection of COVID-19 infections.[16] The details of this model are included in Appendix 1. The learning of Ali-M3 was stopped before our evaluation. We set a cut-off point for the model output at 0.5, because this cut-off point was used during the developing stage. The investigators who entered the CT images data into Ali-M3 were blinded to the RT-PCR results.

Reference standard

The diagnosis of COVID-19 was established by the RT-PCR test, which detected the nucleic acid of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in the sputum, throat swabs, and secretions of the lower respiratory tract samples.[18] We established the RT-PCR tests as the main reference standard. Although the findings of chest CT, interpreted by radiologists, were included as the reference standard in the derivation study, we did not include it as the reference standard in the present study.

Statistical analysis

Statistical analysis was performed using R statistical software, version 3.6.3 (R Foundation for Statistical Computing). Data analysis was performed in a complete-case dataset. Continuous variables are presented as means (standard deviation) and categorical variables are presented as counts and percentages. Using the RT-PCR results as reference, the area under the curve (AUC), sensitivity, specificity, positive-predictive value, and negative-predictive value of the likelihood of COVID-19 as derived from the Ali-M3's analysis of chest CT imaging were calculated. A 95% confidence interval (CI) was determined by the Wilson score method. The goodness-of-fit was calculated using the Le Cessie–Van Houwelingen normal test statistic for the unweighted sum of squared errors.

Sensitivity analysis

1. Moving cut-off point

The objective of this study was to determine whether this AI model could be used as a screening tool for COVID-19 in the real world. In a clinical situation, physicians require an accurate diagnosis of COVID-19; hence, they insist more on sensitivity than on specificity. For sensitivity analysis, we moved the cut-off point and observed sensitivities and specificities to minimize overlooking COVID-19 patients.

2. Simulation of imperfect reference

In the main analysis, we assumed RT-PCR as the perfect reference (100% sensitivity and 100% specificity). However, in the real world, RT-PCR is not the perfect reference since the sensitivity of the RT-PCR test was estimated at 0–80%.[19] To evaluate the effect of this imperfect reference, we calculated the sensitivity, specificity, and AUC of Ali-M3 using the methods and R code described in the Supplemental Material when varying the sensitivity, but fixing the specificity of RT-PCR at 100%.[20]

3. Effect of the number of days after symptom onset

The number of days that have passed since the onset of symptoms affects the performance of antibody and RT-PCR tests in COVID-19 patients.[19, 21] However, it was not clear if this could affect CT images in COVID-19 patients. Sensitivity and specificity were calculated for a group of patients whose symptom onset date was known, among those were those with 14 days or more, as well as those at every 2 days from 0 to 13 days after symptom onset.

4. Effect of symptom severity

Imaging is not routinely indicated as a screening test for COVID-19 in asymptomatic individuals.[22] However, CT images are used in assessment of disease severity. We established the severity by evaluating whether oxygen therapy was required and if the patient was asymptomatic while undergoing CT.

5. Effect of reconstruction slice

The thickness of the reconstruction slice can affect the diagnostic performance.[23] We separated the dataset for the main analysis by a 3-mm reconstruction slice thickness to account for the fissure in our data set between 3 mm and 4 mm and calculated the performance of the model in each dataset.

Results

Study population characteristics

Figure 1 shows the patient flow diagram. Data of 749 patients were evaluated. We assessed 617 symptomatic patients in this validation study. The characteristics of the study population for the main analysis datasets are shown in Table 1. Overall, 289 patients (46.8%) were diagnosed with COVID-19 using the RT-PCR test. Thirteen patients need more than two RT-PCR tests before being diagnosed with COVID-19. Major symptoms were dry cough (37.6%), fever (33.5%), and sore throat (25.8%).

Model performance

The performance of the confidence score after validation among symptomatic patients is shown in Figure 2. The performance of the confidence score was $P = 0.156$ for the goodness-of-fit, and the AUC was 0.797 (95% CI 0.762–0.833). The relationship between the score and predicted probability is shown in Figure 2. The optimal cut-off point with maximal sensitivity and specificity was 0.5, and the sensitivity and specificity were 80.6% (233 of 289) [95% CI: 75.6–85.0%] and 68.3% (224 of 328) [95% CI, 63.0–73.3%], respectively.

Sensitivity analysis

1. Moving cut-off point

   Table 2 shows the relationship between cut-off points for the confidence score and performance. When the cut-off point was 0.2, the sensitivity and specificity were 89.2% and 43.3%, respectively.

2. Simulation of imperfect reference

Figure 3 shows the sensitivity and specificity, with the assumption of imperfect reference of RT-PCR test. The AUC was 0.865. When the cut-off point was set at 0.5, using the Youden Index, the sensitivity and specificity were 80.6% and was 81.3%, respectively. When the cut-off point was set at 0.2, the sensitivity and specificity were 89.2% and 51.9%, respectively.

3. Effect of number of days after symptom onset

Of all symptomatic patients, 600 patients (97.2%) were included in this sensitivity analysis. Of these, the number of days after the onset of symptoms was not known for 17 patients. Figure 4 shows the relationship between test performance and the number of days since the onset of symptoms when the confidence score of Ali-M3 was set at 0.5 or 0.2. Sensitivity values started at 0.7 and increased up to 1.0 until 10–11 days in both cases. However, specificity values remained similar across the strata. The sensitivity increased over 0.9 when the confidence score was set at 0.2 than when the confidence score was set at 0.5.

4. Changing the eligibility criteria

The effects of changing the criteria for patient eligibility are shown n Figure 5.

Dataset focused on asymptomatic patients

There were 86 asymptomatic patients (RT-PCR positive: 37). Using these patients only, the AUC was 0.623. When the cut-off point was 0.5, the sensitivity and specificity were 51.4% and 59.2%, respectively. When the cut-off point was 0.2, the sensitivity and specificity were 44.9% and 73.0%, respectively.

Dataset focused on patients needing oxygen therapy

There were 223 patients who required oxygen support (RT-PCR positive: 97). When using these patients only, the AUC was 0.828. When the cut-off point was set at 0.5, the sensitivity and specificity were 88.7% and 57.9%, respectively. When the cut-off point was set at 0.2, the sensitivity and specificity were 97.9% and 34.9%, respectively.

5. Effect of the thickness of the CT reconstruction slice of CT

   There were 320 patients (RT-PCR positive: 121) with a reconstruction slice thickness of under 3 mm When considering these patients only, the AUC was 0.825. When the cut-off point was set at 0.5, the sensitivity and specificity were 82.6% and 69.7%, respectively. When the cut-off point was set at 0.2, the sensitivity and specificity were 94.2% and 51.5%, respectively. In patients with a reconstruction slice thickness over 3 mm, the AUC was 0.789 (Supplement Figure 1)

Discussion

In this external validation study, our results indicated that Ali-M3 could be useful for early triage of suspected COVID-19 patients with symptoms at a lower cut-off. In particular, higher accuracy was observed in patients with higher severity and a few days since symptom onset, and with images with a thinner reconstructed CT slice thickness.

Currently, all patients with symptoms, such as fever, are triaged as COVID-19 patients. Thus, medical practitioners must use PPE for each patient.[24] Additionally, bed zoning is essential to avoid contamination of non-infected patients.[25] On the other hand, under-triage cause hospital infections through admission of infected patients to facilities. This should continue until a definitive diagnosis is established. Since Ali-M3 is available on the cloud, the physician can receive the results immediately by sending the digital imaging and communications in medicine images from the ordinal picture archiving and communication system. When applying triage, clinicians require sufficient accuracy in terms of sensitivity, but specificity is less important.[19] The high sensitivity obtained at a cut-off of 0.2 with the AI diagnosis is useful for exclude the diagnosis of COVID-19.

Ali-M3 also has the potential to support a diagnosis of COVID-19. The tools currently used for diagnosing COVID-19 infection are antibody, antigen, and RT-PCR tests. Both antigen and RT-PCR tests use tracheal secretions or saliva. An antigen test requires an antigen protein above a given detectable

level, and is currently inferior to RT-PCR tests. As the same patient sample is used, the antigen test cannot support the RT-PCR test. The RT-PCR test is currently used as a gold standard, but the sensitivity changes depending on the number of days after the onset of symptoms.[19] Therefore, for an exclusion diagnosis, multiple tests staggered over time are needed, rather than a single negative RT-PCR test. Even when this test is performed as rapidly as possible, it still requires a few days to obtain multiple test results. On the other hand, Ali-M3 uses the configurational information of patients' lungs, and can add different information than obtained from RT-PCR, thereby complementing the drawbacks of RT-PCR.

In this study, the diagnostic accuracy at the validation stage was lower than the accuracy at the development stage. A two-gate (case–control) design was used in the development of the AI system but in the present study for evaluating the ability of Ali-M3 to assess a COVID-19 diagnosis by chest CT image, we used a single-gate (cohort) design. Although many studies have used the two-gate design in evaluation of AI for the diagnosis of COVID-19,[26] the two-gate design is generally prone to overestimation of diagnostic test results.[27] Thus, blindly using the results based on a two-gate design in a clinical situation can be inappropriate. Moreover, other factors should be considered. With the use of a two-gate design, the fact that RT-PCR is an imperfect reference standard is typically ignored. Furthermore, performing culture and tests to ascertain the true sensitivity of this test is difficult. In the present study, we simulated the diagnostic ability of Ali-M3 with consideration that the sensitivity of the reference standard was imperfect, which leads to underestimation of the specificity and AUC of Ali-M3, without distortion of the sensitivity. Furthermore, the outcomes while developing Ali-M3 and while examining its adequacy were different. Taking into account the patient flow in China, the outcomes at the development stage were set as positive cases with RT-PCR negative results and positive CT image findings.[28] This had a small effect on the sensitivity, but a large effect on the specificity. For example, if in the development stage, 33.9% of the positive patients had negative RT-PCR results and positive CT image findings,[28] then the performance that showed a sensitivity of 98.5% and specificity of 99.2% in the developing Ali-M3,[16] changes from 97.7% to 100% for sensitivity and from 80.8% to 81.6% for specificity when positive RT-

PCR is the only reference used. Upgrading to a diagnostic AI that targets only RT-PCR-positive cases at the development stage is desirable.

This study had some limitations. First, the differentiation performance of Ali-M3 was poor in asymptomatic patients; thus, Ali-M3 should not be used to screen asymptomatic patients. While an alternative to the RT-PCR test for COVID-19 is expected in terms of screening for nosocomial infections and screening on admission for patients with other diseases, Ali-M3 is not recommended for this purpose. Second, we could not differentiate COVID-19 from other viral pneumonias. Compared to the past five seasons, the number of Japanese people infected with influenza during this season was markedly low.[29] In fact, only a few cases in our cohort were diagnosed with other viral pneumonias. Third, it could not reflect the difference in imaging features caused by different COVID-19 types. In addition to type A COVID-19 that was initially prevalent in Asia, type B and type C were prevalent in Europe and in the United States. These different types were not determined in the PCR test, and thus we could not evaluate these differences.

In conclusion, we conducted a retrospective cohort study for external validation of Ali-M3. Our results indicated that AI-based CT diagnosis could be useful for a diagnosis of exclusion of COVID-19 in symptomatic patients, particularly those requiring oxygen and with only a few days since symptom onset. Using Ali-M3 support can reduce PPE consumption and prevent hospital infections through the admission of covertly infected patients. Moreover, Ali-M3 also has the potential to support the diagnosis of RT-PCR for suspected COVID-19 patients. However, as Ali-M3 had some limitations in terms of development, further studies and learning are warranted for updating the system.

References

1. Maves RC, Downar J, Dichter JR, Hick JL, Devereaux A, Geiling JA, Kissoon N, Hupert N, Niven AS, King MA *et al*: **Triage of Scarce Critical Care Resources in COVID-19 An Implementation Guide for Regional Allocation: An Expert Panel Report of the Task Force for Mass Critical Care and the American College of Chest Physicians**. *Chest* 2020, **158**(1):212-225.
2. Carenzo L, Costantini E, Greco M, Barra FL, Rendiniello V, Mainetti M, Bui R, Zanella A, Grasselli G, Lagioia M *et al*: **Hospital surge capacity in a tertiary emergency referral centre during the COVID-19 outbreak in Italy**. *Anaesthesia* 2020, **75**(7):928-934.
3. Li Y, Xia L: **Coronavirus Disease 2019 (COVID-19): Role of Chest CT in Diagnosis and Management**. *AJR American journal of roentgenology* 2020:1-7.
4. Salehi S, Abedi A, Balakrishnan S, Gholamrezanezhad A: **Coronavirus Disease 2019 (COVID-19): A Systematic Review of Imaging Findings in 919 Patients**. *AJR American journal of roentgenology* 2020:1-7.
5. Zhou S, Wang Y, Zhu T, Xia L: **CT Features of Coronavirus Disease 2019 (COVID-19) Pneumonia in 62 Patients in Wuhan, China**. *AJR American journal of roentgenology* 2020:1-8.
6. Chaganti S, Balachandran A, Chabin G, Cohen S, Flohr T, Georgescu B, Grenier P, Grbic S, Liu S, Mellot F *et al*: **Quantification of Tomographic Patterns associated with COVID-19 from Chest CT**. *ArXiv* 2020.
7. Liu K-C, Xu P, Lv W-F, Qiu X-H, Yao J-L, Gu J-F, Wei W: **CT manifestations of coronavirus disease-2019: A retrospective analysis of 73 cases by disease severity**. *European Journal of Radiology* 2020, **126**:108941.
8. Pan F, Ye T, Sun P, Gui S, Liang B, Li L, Zheng D, Wang J, Hesketh RL, Yang L *et al*: **Time Course of Lung Changes at Chest CT during Recovery from Coronavirus Disease 2019 (COVID-19)**. *Radiology* 2020, **295**(3):715-721.
9. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, Pan I, Shi LB, Wang DC, Mei J *et al*: **Performance of Radiologists in Differentiating COVID-19 from Non-COVID-19 Viral Pneumonia at Chest CT**. *Radiology* 2020, **296**(2):E46-E54.
10. Pesapane F, Codari M, Sardanelli F: **Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine**. *Eur Radiol Exp* 2018, **2**(1):35.
11. Huang L, Han R, Ai T, Yu P, Kang H, Tao Q, Xia L: **Serial Quantitative Chest CT Assessment of COVID-19: Deep-Learning Approach**. *Radiology: Cardiothoracic Imaging* 2020, **2**(2):e200075.
12. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q: **Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT**. *Radiology* 2020:200905.
13. Liu W, Liu M, Guo X, Zhang P, Zhang L, Zhang R, Kang H, Zhai Z, Tao X, Wan J *et al*: **Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning**. *European radiology* 2020,

30(6):3567-3575.

14. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M: **Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies**. *BMJ (Clinical research ed)* 2020, **368**:m689.
15. Wynants L, Van Calster B, Bonten MMJ, Collins GS, Debray TPA, De Vos M, Haller MC, Heinze G, Moons KGM, Riley RD *et al*: **Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal**. *BMJ (Clinical research ed)* 2020, **369**:m1328.
16. Academy TAD: **COVID-19 AI Assisted Analysis Based On Chest CT Imaging**. In., vol. 2: The Alibaba DAMO Academy; 2020.
17. Gary S. Collins JBR, Douglas G. Altman, Karel G.M. Moons: **Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement**. *Annals of Internal Medicine* 2015, **162**(1):55-63.
18. Lippi G, Simundic AM, Plebani M: **Potential preanalytical and analytical vulnerabilities in the laboratory diagnosis of coronavirus disease 2019 (COVID-19)**. *Clin Chem Lab Med* 2020.
19. Kucirka LM, Lauer SA, Laeyendecker O, Boon D, Lessler J: **Variation in False-Negative Rate of Reverse Transcriptase Polymerase Chain Reaction-Based SARS-CoV-2 Tests by Time Since Exposure**. *Ann Intern Med* 2020, **173**(4):262-267.
20. Limmathurotsakul D, Turner EL, Wuthiekanun V, Thaipadungpanit J, Suputtamongkol Y, Chierakul W, Smythe LD, Day NPJ, Cooper B, Peacock SJ: **Fool's Gold: Why Imperfect Reference Tests Are Undermining the Evaluation of Novel Diagnostics: A Reevaluation of 5 Diagnostic Tests for Leptospirosis**. *Clinical Infectious Diseases* 2012, **55**(3):322-331.
21. Long QX, Liu BZ, Deng HJ, Wu GC, Deng K, Chen YK, Liao P, Qiu JF, Lin Y, Cai XF *et al*: **Antibody responses to SARS-CoV-2 in patients with COVID-19**. *Nat Med* 2020, **26**(6):845-848.
22. Rubin GD, Ryerson CJ, Haramati LB, Sverzellati N, Kanne JP, Raoof S, Schluger NW, Volpi A, Yim JJ, Martin IBK *et al*: **The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society**. *Radiology* 2020, **296**(1):172-180.
23. He L, Huang Y, Ma Z, Liang C, Liang C, Liu Z: **Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule**. *Sci Rep* 2016, **6**:34921.
24. Organization WH: **Rational use of personal protective equipment (PPE) for coronavirus disease (COVID-19): interim guidance, 19 March 2020**. In.: World Health Organization; 2020.
25. Liu J, Yang J, Li S, Chen J, Yang L, Zhao Z, Hong L: **Gynecological prevention and control model based on ward rearrangement and zoning management in pandemic period of COVID-19**. *Panminerva Med* 2020.
26. Pham TD: **A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks**. *Sci Rep* 2020, **10**(1):16942.
27. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM: **Evidence of bias and**

variation in diagnostic accuracy studies. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2006, **174**(4):469-476.

28. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L: **Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases**. *Radiology* 2020:200642.

29. Sakamoto H, Ishikane M, Ueda P: **Seasonal Influenza Activity During the SARS-CoV-2 Outbreak in Japan**. *JAMA* 2020, **323**(19):1969-1971.

Figure legends

Figure 1. Patient flow.

Abbreviations: CT, computed tomography; RT-PCR, reverse transcription polymerase chain reaction; DICOM, digital imaging and communications in medicine

Figure 2. Differential performance of Ali-M3 for coronavirus disease in symptomatic patients.

(A) A plot of test sensitivity (y-coordinate) versus its false-positive rate (x-coordinate) obtained at each cutoff level confidence score. The area under the receiver operating characteristic curve is 0.797 and the Youden index is 0.50. (B) A plot of test sensitivity, specificity, positive predictive value (PV+), and negative predictive value (PV-) in y-coordinate versus confidence score obtained from Ali-M3 in x-coordinate. The PV+ is dark gray and the PV- is light gray. The maximum PV+ is 46.8% and the maximum PV- is 53.2%. (C) This graph shows the goodness-of-fit. The dashed line is an ideal line that predicts the probability obtained from the confidence score of Ali-M3 equal to the actual probability. The pointed line is the fitted line that is estimated with non-linear assumption alone. The dashed line is the fitted line that is estimated with non-linear assumption and considering the bias in nonparametric estimation using the le Cessie-van Houwelingen method.

Figure 3. Relationship between test performance and the number of days after the onset of symptoms.

(A) The graph shows the relationship between test performance and the number of days after the onset of symptoms when the confidence score from Ali-M3 is at 0.20. (B) The graph shows the relationship between test performance and the number of days after onset of symptoms when the confidence score from Ali-M3 is at 0.50. Light gray bar shows the number of patients included in the strata of days after the onset of symptoms, following the right axis. One stratum includes 2 days from day 0 to day 13. The stratum to the extreme right includes 14 days or more. Following the left axis, solid lines are sensitivity in strata and dash lines are specificity in strata.

Figure 4. Receiver operating characteristic (ROC) curves in ignoring imperfect reference and considering imperfect reference.

(A) A plot of test sensitivity (y-coordinate) versus its false-positive rate (x-coordinate) obtained at each cutoff level confidence score ignoring imperfect reference. The area under the ROC curve is 0.797. (B) A plot of test sensitivity (y-coordinate) versus its false-positive rate (x-coordinate) obtained at each cutoff level confidence score considering imperfect reference. The area under the ROC curve is 0.865.

Figure 5. Differential performance of Ali-M3 for coronavirus disease in asymptomatic patients and patients using oxygen support.

(A) A plot of test sensitivity (y-coordinate) versus its false-positive rate (x-coordinate) obtained at each cutoff level confidence score in asymptomatic patients. The area under the receiver operating characteristic (ROC) curve is 0.623 and the Youden index is 0.25. (B) A plot of test sensitivity, specificity, positive predictive value (PV+), and negative predictive value (PV-) in y-coordinate versus confidence score obtained from Ali-M3 in x coordinate among asymptomatic patients. The PV+ is dark gray and PV- is light gray. The maximum PV+ is 43.0% and maximum PV- is 57.0%. (C) A plot of test sensitivity (y-coordinate) versus its false-positive rate (x-coordinate) obtained at each cutoff level confidence score in patients using oxygen support. The area under the ROC curve is 0.623 and the Youden index is 0.25. (D) A plot of test sensitivity, specificity, PV+, and PV- in y-coordinate versus confidence score obtained from Ali-M3 in x-coordinate in patients using oxygen support. The PV+ is dark gray and the PV- is light gray. The maximum PV+ is 43.5% and the maximum PV- is 56.5%.

Table

Table 1. Demographics of patients' characteristics.

Variable	Symptomatic patients	Patients using oxygen support	Asymptomatic patients
N	617	(223)	(86)
Age (years old) [+]	59.6 (19.2)	68.3 (16.4)	54.5 (22.4)
Sex (Male)	377 (61.2)	158 (70.9)	40 (46.5)
Real-time PCR test (Positive)	289 (46.8)	97 (43.5)	37 (43.0)
Body temperature (≥ 37°)	391 (66.5)	143 (69.8)	
Systolic Blood Pressure (≤ 90 mmHg)	18 (3.2)	11 (5.2)	
Pulse (≥ 120 bpm)	48 (8.2)	22 (10.2)	
Respiratory rate (≥ 25 /minute)	92 (20.5)	64 (38.3)	
Saturation of percutaneous oxygen (≤ 92 %)	105 (17.7)	62 (28.7)	
Oxygen use	223 (36.1)	223 (100.0)	
Vasopressor use	14 (2.3)	14 (6.3)	
Distribution of symptoms reported			
Dry cough	232 (37.6)	67 (30.0)	
Chills	91 (14.7)	40 (17.9)	

Sore throat	159 (25.8)	38 (17.0)	
Diarrhea	66 (10.7)	17 (7.6)	
Joint or muscle pain	46 (7.5)	12 (5.4)	
Conjunctivitis	30 (4.9)	9 (4.0)	
Loss of smell or taste	55 (8.9)	21 (9.4)	
Exposure history			
No	484 (78.4)	191 (85.7)	62 (72.1)
Within family	39 (6.3)	11 (4.9)	6 (7.0)
Other persons	94 (15.2)	21 (9.4)	18 (20.9)
Any international travel	44 (7.1)	6 (2.7)	9 (10.5)
Current Smoking	99 (16.0)	41 (18.4)	11 (12.8)
Past medical history			
Cardiac artery disease	46 (7.5)	24 (10.8)	4 (4.7)
Stroke	60 (9.7)	34 (15.2)	2 (2.3)
Chronic heart failure	69 (11.2)	43 (19.3)	4 (4.7)
Chronic kidney disease	58 (9.4)	33 (14.8)	7 (8.1)
Chronic obstructive pulmonary disease	69 (11.2)	34 (15.2)	7 (8.1)
Malignancy	105 (17.0)	62 (27.8)	8 (9.3)

Immune deficiency	32 (5.2)	17 (7.6)	1 (1.2)
Hypertension	119 (19.3)	71 (31.8)	11 (12.8)
Diabetes	116 (18.8)	64 (28.7)	13 (15.1)
Any other disease	188 (30.5)	73 (32.7)	29 (33.7)

PCR, polymerase chain reaction; bpm, beats per minute

*Patients using oxygen support were included in symptomatic patients

+ is continuous data and the others are count data. Continuous variables are expressed as mean (SD) and count data as number (percentage).

Table 2. Moving cut-off confidence score and test performance.

Confidence score	0.50	0.40	0.30	0.20	0.10
Sensitivity	0.806 (0.755 - 0.850)	0.837 (0.789 - 0.877)	0.854 (0.808 - 0.893)	0.892 (0.851 - 0.925)	0.910 (0.870 - 0.940)
Specificity	0.682 (0.629 - 0.732)	0.612 (0.557 - 0.665)	0.545 (0.490 - 0.600)	0.432 (0.378 - 0.488)	0.375 (0.322 - 0.429)

AUC (95% confidence interval).

Figure1

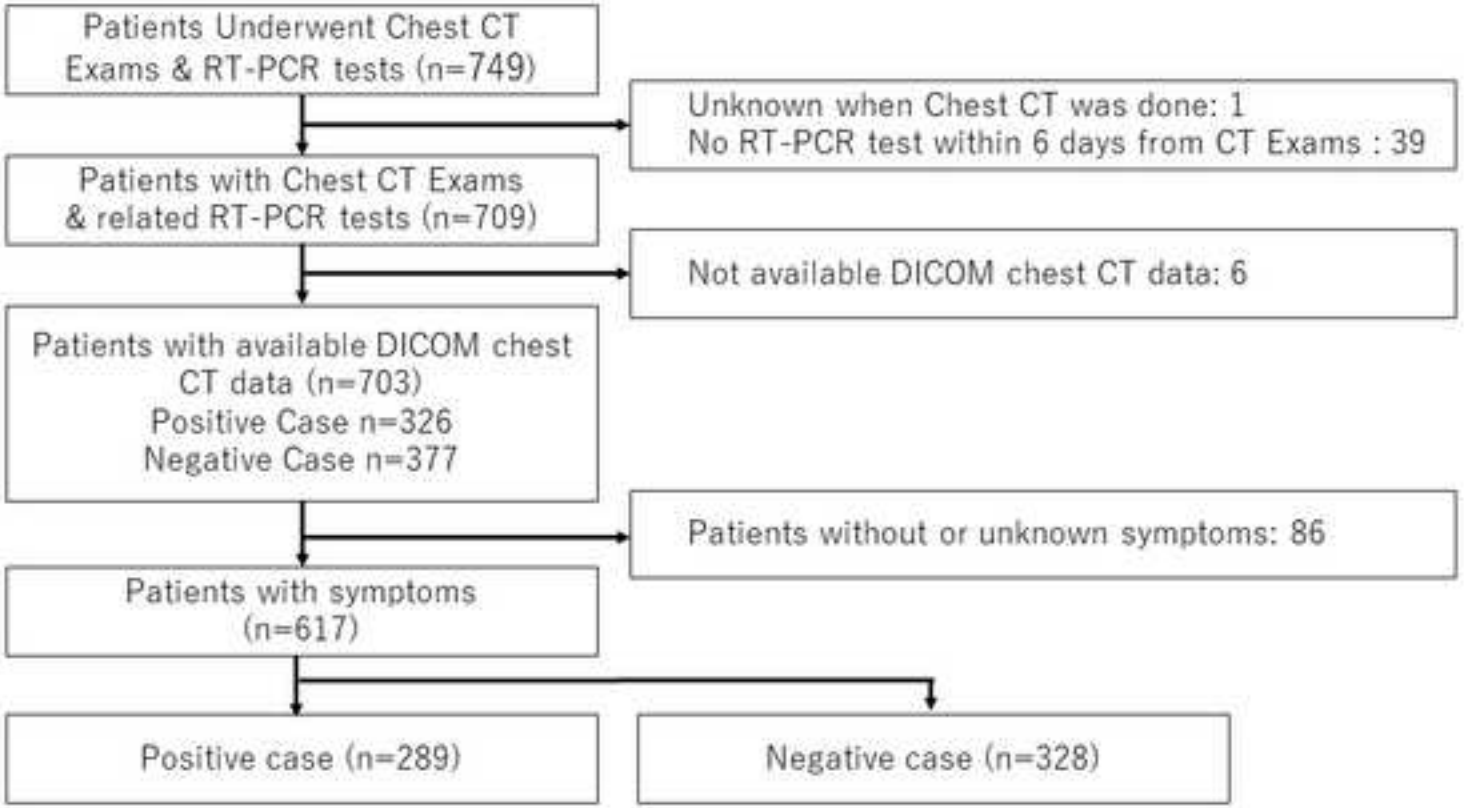


Figure2

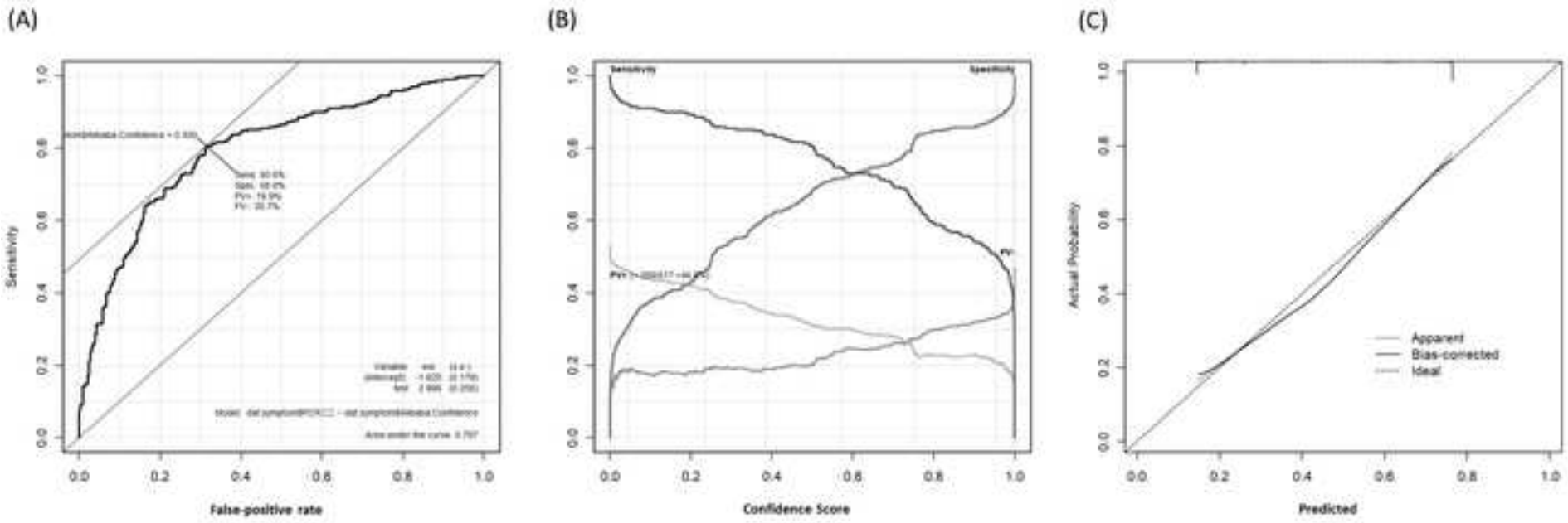
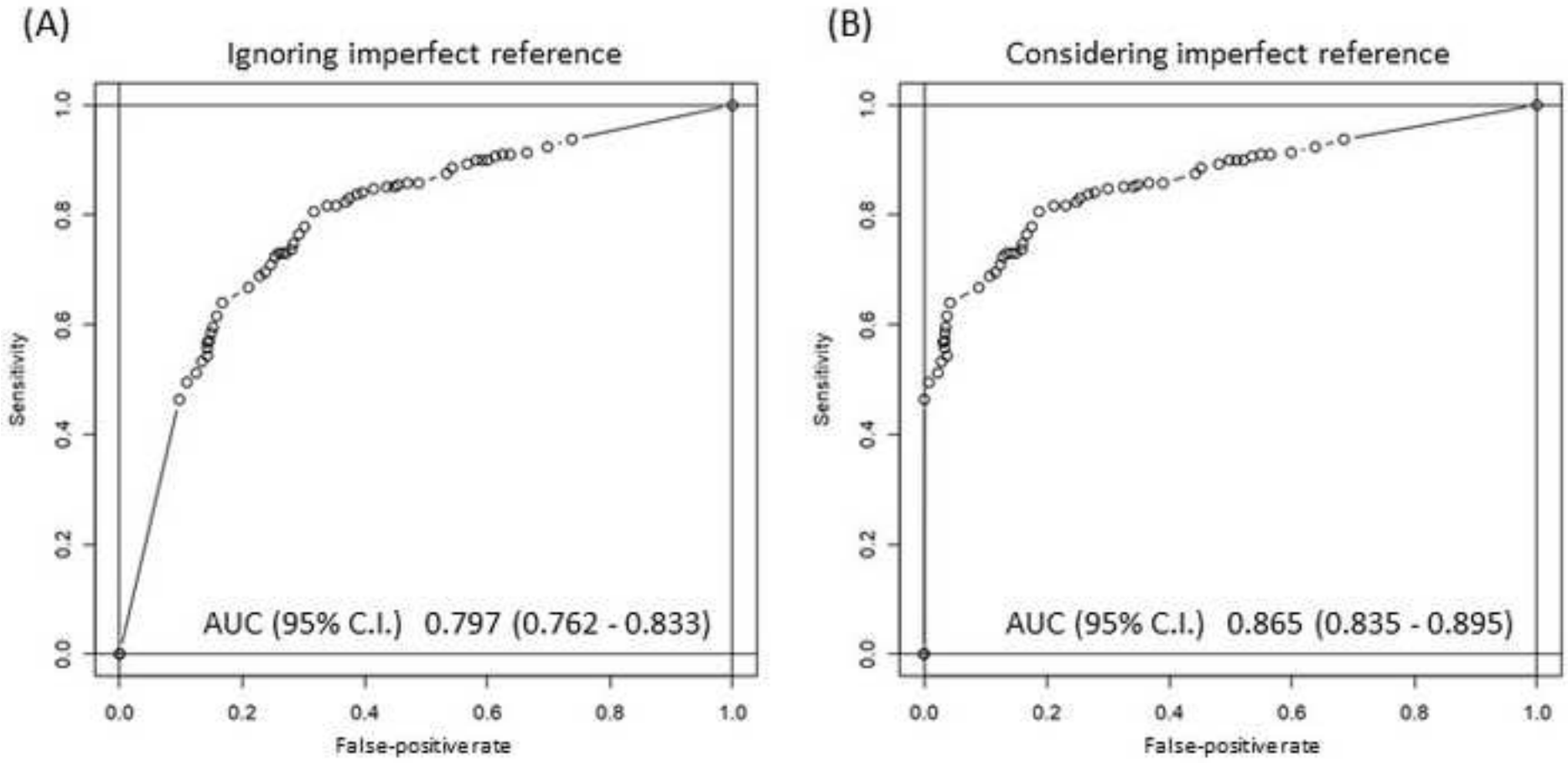


Figure3



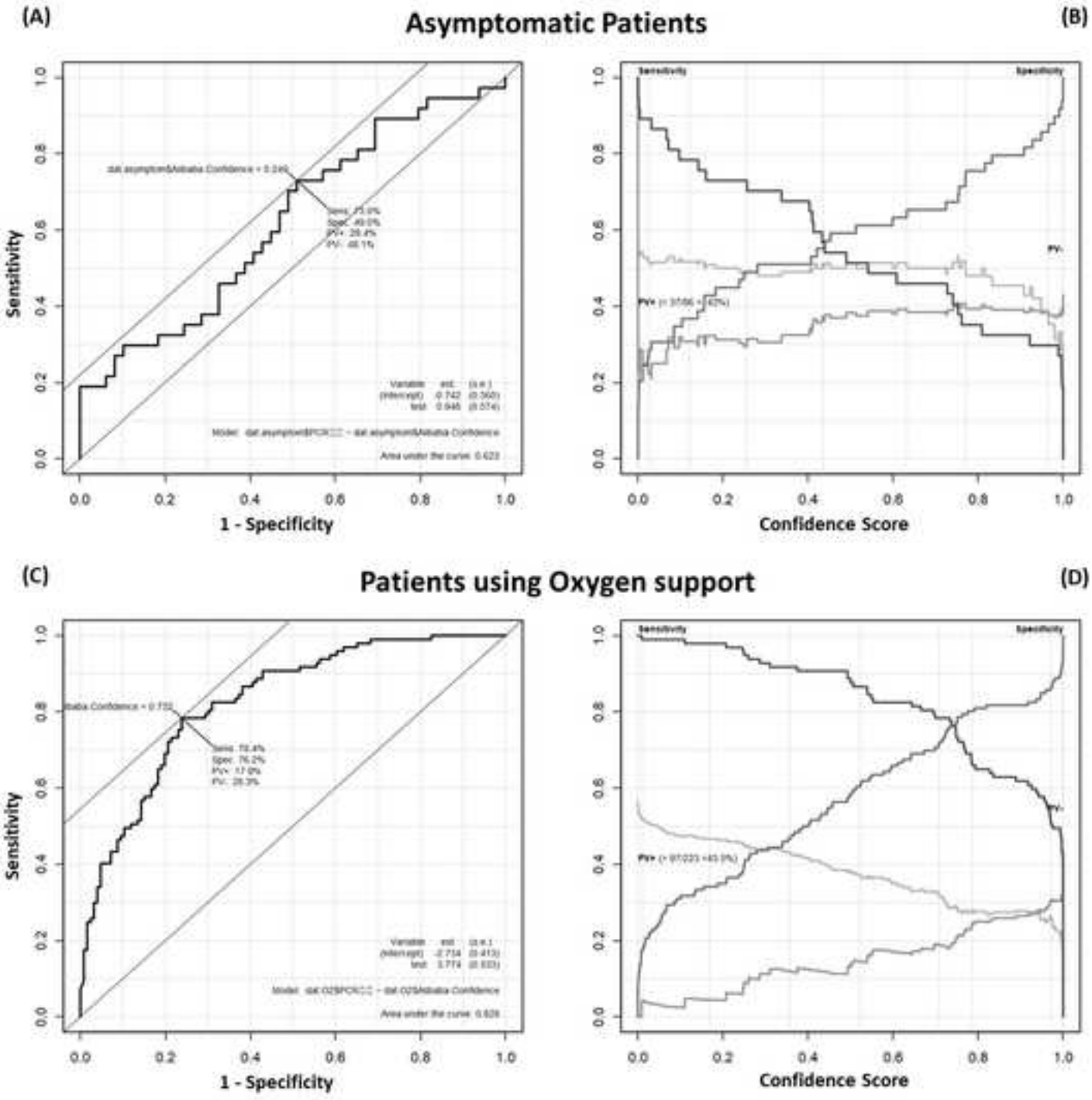
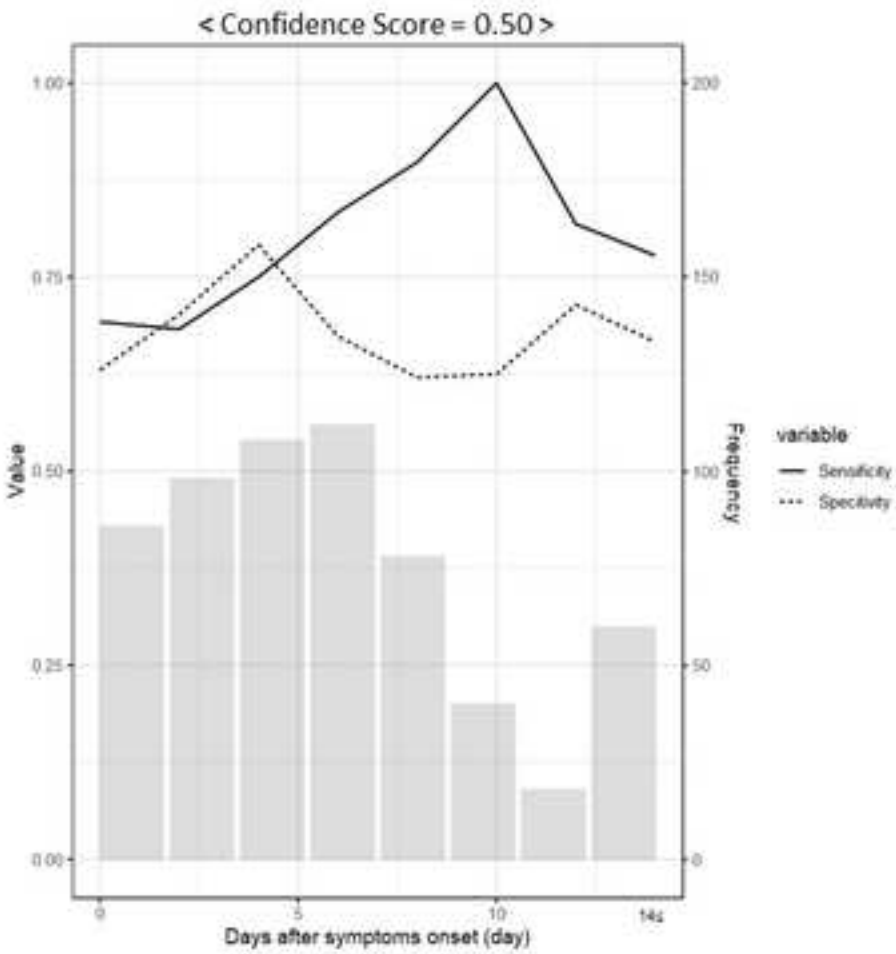
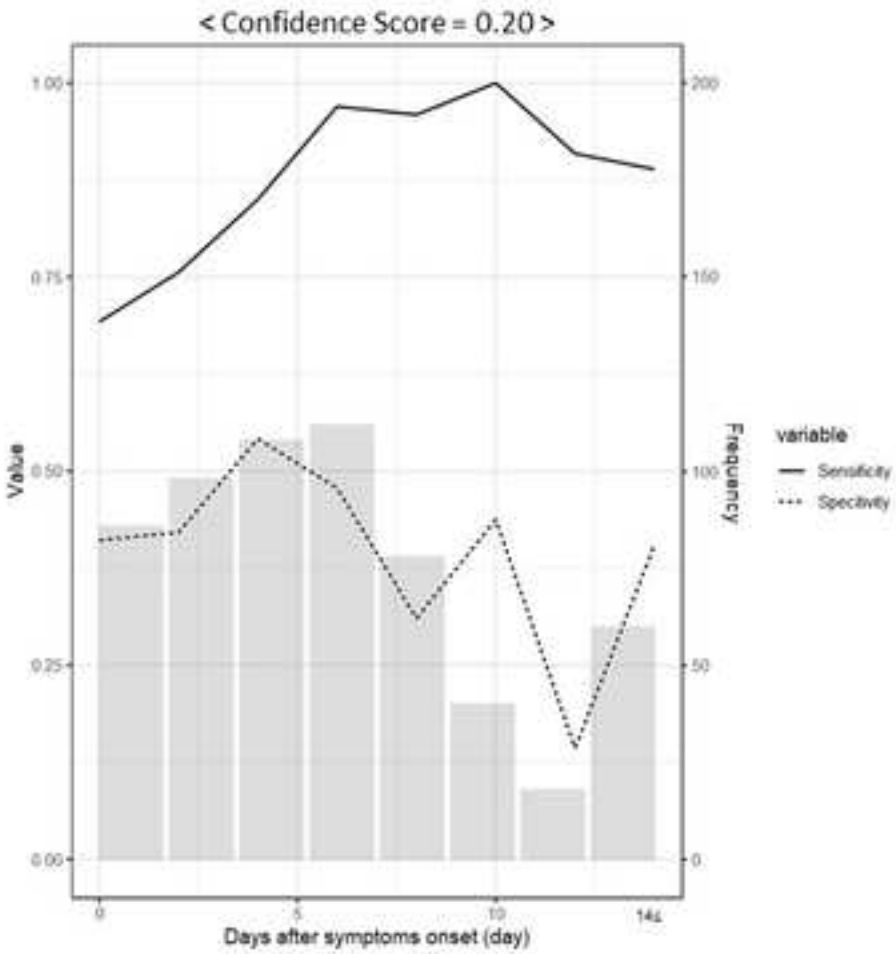


Figure5





Compliance with ethical standards:

Guarantor:

The scientific guarantor of this publication is Tatsuyoshi Ikenoue.

Conflict of interest:

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Funding:

The authors state that this work has not received any funding.

Statistics and biometry:

No complex statistical methods were necessary for this paper.

Informed consent:

Written informed consent was waived by the Institutional Review Board.

Ethical approval:

Institutional Review Board approval was obtained.

Study subjects or cohorts overlap:

Any study subjects or cohorts have not been previously reported.

Methodology:

- retrospective
- diagnostic or prognostic study
- multicentre study

Acknowledgments

We thank M3 Inc., and Clinical Porter for the support with providing free Ali-M3 and data storage, although they did not participate in the preparation protocol and manuscript. To want to access Ali-M3, reader can contact M3 (m3-ai-lab@m3.com). We also thank Ms. Kyoko Wasai, who assisted retrieving data.