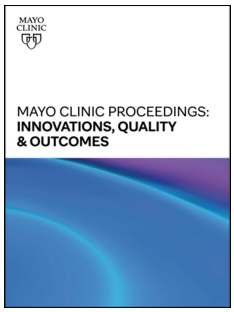
# Journal Pre-proof




Evaluating the Performance of a Commercially Available AI Algorithm for Automated Detection of Pulmonary Embolism on CECT and CTPA of COVID-19 Patients

Karim A Zaazoue, MD, Mathew R McCann, MD, Ahmed K Ahmed, MD, MSc, Isabel O Cortopassi, MD, Young M. Erben, MD, Brent P Little, MD, Justin T Stowell, MD, Beau B Toskich, MD, Charles A Ritchie, MD

PII: S2542-4548(23)00011-5

DOI: https://doi.org/10.1016/j.mayocpiqo.2023.03.001

Reference: PIQO 474

To appear in: *Mayo Clinic Proceedings: Innovations, Quality & Outcomes*

Received Date: 6 January 2023

Accepted Date: 3 March 2023

Please cite this article as: Zaazoue KA, McCann MR, Ahmed AK, Cortopassi IO, Erben YM, Little BP, Stowell JT, Toskich BB, Ritchie CA, Evaluating the Performance of a Commercially Available AI Algorithm for Automated Detection of Pulmonary Embolism on CECT and CTPA of COVID-19 Patients, *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* (2023), doi: https://doi.org/10.1016/j.mayocpiqo.2023.03.001.

# Evaluating the Performance of a Commercially Available AI Algorithm for Automated Detection of Pulmonary Embolism on CECT and CTPA of COVID-19 Patients

Karim A Zaazoue[1], MD, Mathew R McCann[2], MD, Ahmed K Ahmed, MD, MSc[1], Isabel O Cortopassi[2], MD, Young M. Erben[3], MD, Brent P Little[2], MD, Justin T Stowell[2], MD, Beau B Toskich[1], MD, Charles A Ritchie[1], MD

(1) Division of Interventional Radiology, Department of Radiology, Mayo Clinic, Jacksonville, Florida

(2) Division of Cardiothoracic Imaging, Department of Radiology, Mayo Clinic, Jacksonville, Florida

(3) Department of Vascular Surgery, Mayo Clinic, Jacksonville, Florida

All authors declare no conflict of interest

Address for correspondence: Charles A. Ritchie, 4500 San Pablo Road S, Jacksonville, FL 32224, 1-904-953-2000, Ritchie.charles@mayo.edu

Word Count: 2,251

Tables/Figures: 9

**Key Results**

- Sensitivity and specificity of the artificial intelligence algorithm for the detection of pulmonary embolism in hospitalized COVID-19 patients was 93.2% and 99.6%, respectively.
- The degree of parenchymal disease characterized by the total severity scoring system did not affect the accuracy of AI (P=.375).
- The accuracy of artificial intelligence was affected by the mean attenuation in the pulmonary vasculature and was significantly higher in the CTPA group compared to the contrast-enhanced CT (P<.001) with an optimal cutoff value for AI at 362 HU (P=.048).

**Objective**

To investigate the performance of a commercially available artificial intelligence (AI) algorithm for detection of pulmonary embolism (PE) on contrast-enhanced CTs in patients hospitalized for COVID-19.

**Patients & Methods:**

Retrospective analysis was performed of all contrast-enhanced chest CTs on patients admitted for COVID-19 between March 2020 and December 2021. Based on the original radiology reports, all PE-positive exams were included (n=527). Using a reversed flow single gate diagnostic accuracy case-control model, a randomly selected cohort of PE-negative exams (n=977) was included. Pulmonary parenchymal disease severity was assessed for all included studies using a semi-quantitative system, the Total Severity Score (TSS). All included CTs were sent for interpretation by the commercially available AI algorithm, Aidoc. Discrepancies between AI and original radiology reports were resolved by three blinded radiologists, who rendered a final determination of indeterminate, positive, or negative.

**Results**

A total of 78 studies were found to be discrepant, of which 13 (16.6%) were deemed indeterminate by readers and excluded. The sensitivity and specificity of AI was 93.2%; (95% confidence interval [CI] 90.6-95.2%), and 99.6%; (95% CI 98.9-99.9%), respectively. AI's accuracy for all TSS groups (mild, moderate, severe) was high (98.4%, 96.7%, and 97.2%, respectively). AI was more accurate in PE detection on CTPAs vs CECTs ($P < .001$), with optimal HU of 362 ($P=.048$).

**Conclusion**

The AI algorithm demonstrated high sensitivity, specificity, and accuracy for PE on contrast enhanced CTs in COVID-19 patients regardless of parenchymal disease. Accuracy was significantly affected by the mean attenuation of the pulmonary vasculature. How this affects the legitimacy of the binary outcomes reported by AI is not yet known.

**Abbreviations:**

AI = artificial intelligence

CECT = contrast-enhanced computed tomography

COVID-19 = Coronavirus disease 2019

CTPA = computed tomography pulmonary angiogram

FN = false negative

FP = false positive

HU = Hounsfield unit

PE = pulmonary embolism

SARS-CoV-2= severe acute respiratory syndrome coronavirus 2

TN = true negative

TP = true positive

## Introduction

Artificial intelligence (AI) encompasses a broad field of research and engineering geared towards the creation of "intelligent machines". Various AI applications have been employed in the field of radiology over the years, such as computer-assisted detection (CAD) technology beginning in the 1980s and, more recently, deep learning techniques to assist in image interpretation.[1] In previous studies, AI has demonstrated high sensitivity and specificity for automated detection of pulmonary embolism (PE) in patients undergoing computed tomography pulmonary angiogram (CTPA) and contrast-enhanced computed tomography (CECT).[2-5]

Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has been associated with prothrombotic states due to multiple factors such as anti-cardiolipin and antibeta2-glycoprotein I leading to deep venous thrombosis, PE, and arterial thrombosis.[6-8] Recent publications have demonstrated that the risk of developing PE in COVID-19 patients ranged from 5.7 to 24 percent and has been reported as high as 29 percent in ICU patients.[9-11] The aim of this study was to evaluate the performance of a commercially available AI algorithm in detecting PE in patients hospitalized for COVID-19 and to ascertain whether the associated parenchymal disease affects the performance of the AI's algorithm, which was trained on patients prior to the global pandemic.

## METHODS

### *Patient selection*

This was an IRB-approved study utilizing a prospectively maintained database of patients hospitalized for COVID-19 in a multisite healthcare system (IRB 20-003457). A retrospective

analysis was performed to identify patients who underwent contrast-enhanced chest CT between March 2020 and December 2021. The medical charts of eligible patients were manually reviewed to confirm that hospitalization was for the treatment of COVID-19. Infections were diagnosed by polymerase chain reaction testing or serology.

Based on the original radiology reports, all PE-positive examinations were identified and included. Using a reversed flow single gate diagnostic accuracy case-control model, a random selection of PE-negative exams was included.

*Imaging Parameters*

Cross-sectional imaging was performed using 64, 128, 256, or 384 row multidetector CT scanners. CTPAs (i.e., examinations specifically tailored for assessment of the pulmonary arteries) were performed utilizing 0.6 mm collimation, 1.5 to 10 mm reconstruction, rotation time 0.25 to 0.5 seconds, and 3 mm/rotation table speed. Up to 100 mL intravenous isoosmotic iodinated contrast agent was used in each study with bolus tracking threshold of 100 Hounsfield Units at the main pulmonary artery.

CECTs (i.e., contrast-enhanced examinations not specifically tailored for assessment of the pulmonary arteries) were performed utilizing a 0.6 mm collimation, 1.5 to 10 mm reconstruction, rotation time of 0.5 seconds, and 3 mm/rotation table speed using a 40 second delay and 75 mL of intravenous isoosmotic iodinated contrast agent. Mean attenuation in Hounsfield units at the pulmonary artery bifurcation was calculated for each included examination.

*Artificial Intelligence*

All CT data was collected retrospectively using the institution's picture archiving and communication system (Visage, Pro Medicus Limited, Australia), anonymized, and submitted for analysis by an AI algorithm (Aidoc, Tel Aviv, Israel). Aidoc provides multiple cloud-based clinical radiology algorithms for computer-aided triage and worklist prioritization. The product uses a proprietary deep convolutional neural network that was trained on tens of thousands of studies acquired on a diverse range of scanners from multiple medical centers around the world prior to the global COVID-19 pandemic.

*Outcome Assessment*

For each CT examination, AI determined a binary classification for PE (i.e., positive, or negative); this result was then compared with the original clinical radiology report. When discrepancies were found between the AI result and the original clinical report, two cardiothoracic radiologists blinded to clinical and AI interpretations and clinical data (JTS and BPL, 5-12 years of experience after completion of training) independently reviewed the studies and rendered a final interpretation of indeterminate, positive, or negative for PE. Disagreements between the two primary radiologists were resolved by a tie-breaking interpretation by a third cardiothoracic radiologist (IOC, 14 years of experience).

Based on the original clinical radiology report and the blinded radiologists' interpretations, the AI results were classified as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). In TP cases, AI identified a thrombus, and the finding was corroborated by the original report and the readers. In TN cases, AI determined the absence of a thrombus, which was confirmed by the original reports and the readers. FPs were defined as any instance in which

AI misinterpreted a finding as a thrombus, which was not identified by the original report, or the readers FN was defined as any instance in which AI did not identify a PE that was determined to be present by the original report and the readers.

*Radiographic Assessment of COVID-19 airspace severity*

The degree of pulmonary parenchymal disease on each CT examination was determined using a semi-quantitative scoring system, the Total Severity Score (TSS). The score was determined by a fourth board-certified radiologist (CAR, 10 years of experience) who was not blinded to the PE status of the studies. Each pulmonary lobe was assigned a score ranging from zero to five, with zero corresponding to no parenchymal involvement and five corresponding to more than 75% involvement; the individual lobe scores were then totalled to yield the TSS.[12] Based on the TSS, studies were classified into three groups: mild (0-8), moderate (9-16), and severe (17-25). A comparison was made between the positive and negative PE studies to ensure TSS was controlled, and randomization was met.

*Statistical methods*

Data was analyzed using IBM's Statistical Package for the Social Sciences (SPSS®, version 28.0). The study population's demographic and clinical information were characterized using descriptive statistics. All parameters analyzed were taken from patients' charts. The Chi-square and ($\chi 2$) Mc Nemar’s test were used to compare categorical variables, and the Mann-Whitney U test was used to investigate the relationship between continuous variables between the two groups. A p-value of $< 0.05$ was taken to be statistically significant. Heterogeneity of effect was assessed via stratified risk analysis.

The performance of the AI algorithm was evaluated by calculating the sensitivity, specificity, and accuracy. 95% Confidence intervals (CI) for sensitivity, specificity, and accuracy are exact Clopper-Pearson confidence intervals.[13]

## Results

*Study design*

A total of 9488 patients were admitted with COVID-19 during the period between March 2020 and December 2021. All positive PE studies were included (N=527) and 977 studies negative for PE were selected as matched controls as outlined earlier (Figure 1).

*Population demographics*

Of the selected patients, 710 (61%) were male, and the median age of the group was 64 years (interquartile range (IQR) 53-74); and the median BMI was 30.2 $kg/m^2$ (IQR 26-35.7). There were 406 (34.8%) patients in ICU care and 233 (20%) on mechanical ventilation at the time of CT examination. 348 (30%) had a smoking history.

*TSS*

The mean TSS for the positive PE and negative PE groups were 15 and 16, respectively. TSS evaluation showed 309 studies (20.7%) being classified as mild, 426 (28.6%) as moderate, and 756 (50.7%) as severe (Figure 2).

*Blinded review*

According to the original clinical radiology reports, 527 were positive and 977 were negative, and after review, a total of 78 studies were discrepant. Thirteen studies were deemed indeterminate by the blinded cardiothoracic radiologists and were excluded from diagnostic accuracy analysis (Table 2). After review and exclusion, there were 526 studies positive for PE and 965 studies negative for PE. The original radiologists demonstrated 16 missed PEs (FN), and 9 incorrectly identified PEs (FP), when compared to the blinded radiologists' review as the reference standard (Figure 3).

*Diagnostic accuracy values*

AI correctly identified 490/526 studies positive for PE and 961/965 studies negative for PE showing a sensitivity and specificity of 93.2%; (95% confidence interval [CI] 90.6-95.2%), and 99.6%; (95% CI 98.9-99.9%), respectively (Table 3) (Figures 4 & 5). AI's accuracy in the mild, moderate, and severe TSS groups was high (98.4%, 96.7%, and 97.2%, respectively) with p-value of .375. AI accuracy was not significantly affected by BMI (AUC = 0.584, P=.069).

*PE anatomy*

For examinations with multiple thrombi, PEs were classified according to the most proximal location in the pulmonary artery tree. There were 107/526 (20.3 %) central (saddle or main pulmonary artery or lobar), 232/526 (44.1%) segmental, and 187/526 (35.6%) subsegmental emboli. AI detected 92.5% of the central PEs, 90.9% of the segmental PEs, and 96.3% of the subsegmental PEs without statistically significant difference between the different locations (P= .097).

*CTPA vs CECT*

There were 1,392 (93.4%) CTPAs and 99 (6.6%) CECTs performed. The accuracy of AI in detecting PE was higher with CTPA compared to CECT (97.8% vs 90.9%, P<.001). For CTPA, the median average region of interest (ROI) attenuation at the pulmonary artery bifurcation was 392 HU (IQR 320-481) compared to 210 HU (IQR 163-267) in the CECT group (P <.001). Receiver operating characteristic (ROC) curve and Youden index analysis showed that the optimal cutoff point for AI's accuracy in detecting PE in this cohort was 362 HU (Area under the curve [AUC]= **0.592** (0.499 -0.685), P=**.048** (figure 6).

**Discussion:**

The present study was designed to evaluate the performance of a commercially available AI algorithm, Aidoc, for automated detection of PE using contrast enhanced chest CT studies in patients hospitalized with COVID-19. Aidoc was created and trained prior to the global pandemic, and we therefore aimed to determine whether infection-related parenchymal disease affected the performance of the algorithm. To our knowledge this has not been previously reported. In this study, the sensitivity and specificity of Aidoc was 93.2% and 99.6%, respectively. The accuracy of the algorithm was not affected by the degree of COVID-19 parenchymal disease as the TSS was matched in both positive and negative PE studies. Additionally, BMI and thrombus location did not affect accuracy.

Overall, sensitivity and specificity for this cohort were similar to those of previously published studies reporting Aidoc's performance in non-COVID-19 patients (sensitivity: 93.2% vs 92.7%, specificity: 99.6% vs 95.5%) (2). This study was not specifically designed to evaluate the

interobserver agreement of the original readers' reports, however, it is pertinent to highlight that there were discrepancies discovered. Identified first by the AI then validated by the blinded readers, there were 16 FN (1.1%) and 9 FP (0.6%). This percentage of FN and FP incidence is similar to previously reported findings by Kligerman et al in which the FN and FP was 1.4 % and 0.74%, respectively. These findings were in a study in which an AI algorithm was not used.[14]

Forty discrepant cases were determined to be inaccurate interpretations by Aidoc, 36 FN (2.4%), and 4 FP (0.3%). However, AI's accuracy was found to be diminished when comparing the CT technique performed (CTPA compared to CECT, 97.8% vs 90.9% respectively, P<.001). Using a stratified risk analysis, the variation in accuracy was related to the difference in pulmonary vascular attenuation between the CTPA and CECT studies (392 HU vs 210 HU, P<.001). The value with the highest Youden index (J) was measured at 362 HU in this cohort. Thus, for the AI to produce the most accurate results, the highest attenuation at the pulmonary artery bifurcation should be above 362 HU, P=.048. This differs from the consensus among radiologists that PE cannot be ruled out if the density is less than 250 HU.[15-16] This would suggest that the Aidoc algorithm may require higher density within the pulmonary vasculature for accurate detection of PE than radiologists, a finding that warrants further investigation.

The blinded readers believed 13/78 discrepant cases were indeterminate and not sufficient to comment on the presence or absence of PE. The most common reason for this classification was motion artifact, and these cases were excluded from the overall calculation of accuracy. The distinction of a study to be indeterminate is a uniquely human capability, at present. Indeterminate or non-diagnostic examinations are described in previous publications that

evaluated human interobserver agreement for the detection of PE on CT and have not been previously discussed in the AI literature previously to our knowledge.[17-20] Aidoc and other algorithms which produce binary outcomes are not capable of making this distinction. Alternative reporting outcomes of different AI algorithms can include percent probability or continuous variable outcomes but are still incapable of excluding studies for any reason. This can negatively affect radiologists who could struggle with believing their own intuition over the algorithm and may increase physician error and liability according to a recent study by Tobia et al. Through an experimental study, the authors looked at liability in incidents of patient harm and found that jurors were more likely to side with AI.[21]

There are several limitations in the design of this study. The retrospective nature and use of the original radiologist reports for screening of disease excludes the possibility for a head-to-head comparison between AI and radiologists. This resulted in only the discrepancies between AI and the original reports being sent to our blinded readers for reinterpretation. Additionally, positive, and negative predictive values could not be calculated secondary to the diagnostic case control design that was used and thus the prevalence of PE in our cohort was higher than the general population. Finally, the random selection of presumed negative studies, for the purposes of matching TSS, likely resulted in an under reporting of FN studies by the original radiologist.

**Conclusion:**

In conclusion, AI exhibits a high sensitivity, specificity, and accuracy in the automated detection of PE in the setting of COVID-19 parenchymal disease. AI performed significantly better at this task on CTPAs compared to CECT ($P<.001$) and the highest accuracy was seen when the

attenuation at the pulmonary artery was more than 362 HU, a finding that warrants further investigation.

## References


1. Doi K. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics. 2007;31(4-5):198-211.
2. Weikert T, Winkel DJ, Bremerich J, Stieltjes B, Parmar V, Sauter AW, et al. Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. European Radiology. 2020;30(12):6545-53.
3. Tajbakhsh N, Shin JY, Gotway MB, Liang J. Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and discriminative image representation. Medical Image Analysis. 2019;58:101541.
4. Huang S-C, Kothari T, Banerjee I, et al. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. npj Digital Medicine. 2020;3(1).
5. Liu W, Liu M, Guo X, et al. Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning. European Radiology. 2020;30(6):3567-75.
6. Abou-Ismail MY, Diamond A, Kapoor S, Arafah Y, Nayak L. The hypercoagulable state in COVID-19: Incidence, pathophysiology, and management. Thrombosis Research. 2020;194:101-15.
7. Helms J, Tacquard C, Severac F, et al. High risk of thrombosis in patients with severe SARS-CoV-2 infection: a multicenter prospective cohort study. Intensive Care Medicine. 2020;46(6):1089-98.
8. Zhang Y, Xiao M, Zhang S, et al. Coagulopathy and Antiphospholipid Antibodies in Patients with Covid-19. New England Journal of Medicine. 2020;382(17):e38.
9. Riyahi S, Dev H, Behzadi A, Kim J, Attari H, Raza SI, et al. Pulmonary Embolism in Hospitalized Patients with COVID-19: A Multicenter Study. Radiology. 2021;301(3):E426-E33.
10. Erben Y, Franco-Mesa C, Gloviczki P, et al. Deep vein thrombosis and pulmonary embolism among hospitalized coronavirus disease 2019–positive patients predicted for higher mortality and prolonged intensive care unit and hospital stays in a multisite healthcare system. Journal of Vascular Surgery: Venous and Lymphatic Disorders. 2021;9(6):1361-70.e1.
11. Ritchie CA, Johnson MM, Stowell JT, et al. Resolution of acute pulmonary embolism using anticoagulation therapy alone in coronavirus disease 2019. J Vasc Surg Venous Lymphat Disord. 2022.

12. Pan F, Ye T, Sun P, et al. Time Course of Lung Changes at Chest CT during Recovery from Coronavirus Disease 2019 (COVID-19). Radiology. 2020;295(3):715-21.
13. CLOPPER CJ, PEARSON ES. THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL. Biometrika. 1934;26(4):404-13.
14. Kligerman SJ, Mitchell JW, Sechrist JW, et al. Radiologist Performance in the Detection of Pulmonary Embolism: Features that Favor Correct Interpretation and Risk Factors for Errors. J Thorac Imaging. 2018;33(6):350-7.
15. Moore AJE, Wachsmann J, Chamarthy MR, et al. Imaging of acute pulmonary embolism: an update. Cardiovascular Diagnosis and Therapy. 2018;8(3):225-43.
16. Chen M, Mattar G, Abdulkarim JA. Computed tomography pulmonary angiography using a 20% reduction in contrast medium dose delivered in a multiphasic injection. World Journal of Radiology. 2017;9(3):143.
17. Wildman-Tobriner B, Ngo L, Mammarappallil JG, et al. Missed Incidental Pulmonary Embolism: Harnessing Artificial Intelligence to Assess Prevalence and Improve Quality Improvement Opportunities. Journal of the American College of Radiology. 2021;18(7):992-9.
18. Joshi R, Wu K, Kaicker J, Choudur H. Reliability of on-call radiology residents' interpretation of 64-slice CT pulmonary angiography for the detection of pulmonary embolism. Acta Radiol. 2014;55(6):682-90.
19. Shaham D, Heffez R, Bogot NR, Libson E, Brezis M. CT pulmonary angiography for the detection of pulmonary embolism: interobserver agreement between on-call radiology residents and specialists (CTPA interobserver agreement). Clin Imaging. 2006;30(4):266-70.
20. Courtney DM, Miller C, Smithline H, et al. Prospective multicenter assessment of interobserver agreement for radiologist interpretation of multidetector computerized tomographic angiography for pulmonary embolism. J Thromb Haemost. 2010;8(3):533-9.
21. Tobia K, Nielsen A, Stremitzer A. When Does Physician Use of AI Increase Liability? Journal of Nuclear Medicine. 2021;62(1):17-21.

**Table 1** Patients' demographics and clinical data (N=1166)

| Variable | | N (%) | Median age (IQR) | Median BMI (IQR) |
|---|---|---|---|---|
| Age | | 1166 (100%) | 64 (53 - 74) years | 30.2 (26 – 35.7) |
| Gender | Male | 710 (61%) | 64 (54.75 - 74) years | **29.5 (26 – 34.5)** |
| | Female | 456 (39%) | 63 (50 - 74) years | **31.6 (26.6 – 37.4)** |
| Smoking | Yes | 348 (30%) | **66 (58 -75)** | **29.5 (25.7 – 34.6)** |
| | No | 818 (70%) | **62 (51 - 73)** | **30.6 (26.3 - 36)** |
| Vaccinated | Yes | 898 (77%) | 64 (55 – 73.7) | 30.8 (26.2 – 35.8) |
| | No | 268 (23%) | 63 (52 - 74) | 30 (26 – 35.6) |
| Intubated | Yes | 933 (80%) | **61 (51 - 71)** | 30 (26.5 – 35.6) |
| | No | 233 (20%) | **65 (53 - 75)** | 30 (25.8 – 35.7) |
| ICU stay | Yes | 759 (65.1%) | 63 (53 - 74) | 30 (26 - 36) |
| | No | 406 (34.8%) | 64 (52 - 74) | 30 (26 – 35.4) |
| Dead | Yes | 208 (17.8%) | **73 (62 - 80)** | **28.8 (25.3 – 34.3)** |
| | No | 958 (82.2%) | **62 (51 - 73)** | **30.5 (26.3 - 36)** |

Bold entries represent statistically significant associations/differences (Mann-Whitney Test)

BMI=body mass index, ICU=intensive care unit, IQR=interquartile range

**Table 2** Artificial Intelligence vs original reporting and vs final review

| | | Original report | | Total | Final review | | Total* |
|---|---|---|---|---|---|---|---|
| | | Positive | Negative | | Positive | Negative | |
| AI | Positive | 474 | 25 | 499 | 490 | 4 | 494 |
| | Negative | 53 | 951 | 1005 | 36 | 961 | 997 |
| Total | | 527 | 977 | 1504 | 526 | 965 | 1491* |

*Indeterminate results excluded

**Table 3** Artificial Intelligence diagnostic evaluation estimates based on the original reporting and after final review

| Reference test | Sensitivity (95% CI) * | Specificity (95% CI) * | Accuracy (95% CI) * |
|---|---|---|---|
| Based on the original reports | 89.94% (89.0 -92.4) % | 97.44% (96.2 - 98.3) % | 94.8% (93.6 - 95.9) % |
| After review of the discrepant cases | 93.2% (90.6 - 95.2) % | 99.6% (98.9 - 99.9) % | 97.3% (96.4 - 98.1) % |

PPV = positive predictive value, NPV = Negative predicative value, CI = confidence interval

*Confidence intervals (CI) for sensitivity, specificity and accuracy are "exact" Clopper-Pearson confidence intervals.

**Figure 1.**

Reversed flow diagnostic case-control design (single gate) flow chart

* Patient was excluded because they underwent lung surgery excluding TSS calculation

**Figure 2.**

Axial CTPA images illustrating the total severity scoring (TSS) semiquantitative assessment. (A) TSS = 0 (mild), (B) TSS = 13 (moderate), and (C) TSS = 25 (severe).

**Figure 3**.

CT showing original reports of two false positive studies. (A) Axial CTPA image shows a soft tissue density misinterpreted as an eccentric filling defect (arrow). (B) The coronal-oblique multiplanar reconstructed image shows presumed filling defect (arrow) between the branch of the right superior pulmonary artery (star) and one of the right superior pulmonary vein tributaries (cross). (C) Coronal reconstructed CTPA image shows soft tissue density (arrow) misinterpreted as a filling defect within the right subsegmental upper lobar artery. (D) Oblique-oblique multiplanar reconstructed CTPA image shows the subsegmental branch (arrowheads) to be patent.

CTPA = Computerized tomography pulmonary angiography

**Figure 4**.

CTPA shows AI's false positive result. (A) AI-generated image highlighting the misinterpreted filling defect within the right middle lobar branch, also seen on (B) the corresponding axial

CTPA image. Note that the presumed filling defect (arrow) is seen inferior to the vessel on both (C) the coronal and (D) sagittal reconstructed images.

AI = Artificial Intelligence, CTPA = Computerized tomography pulmonary angiography

**Figure 5**.

CTPA shows AI's false positive result. (A) AI-generated image highlighting the misinterpreted filling defect, also seen on (B) the corresponding axial CTPA image. (C) Coronal reconstructed image shows the presumed filling defect soft (arrow) between the apical branch of the right superior pulmonary artery (star) and one of the right superior pulmonary vein tributaries (cross).

AI = Artificial Intelligence, CTPA = Computerized tomography pulmonary angiography

**Figure 6**.

The figure shows receiver operating characteristic (ROC) curve for evaluation of highest mean ROI density at PA bifurcation in classifying false/true AI results. Area under the curve (AUC) = **0.592** (0.499 -0.685), p value = **0.048**. Value with the highest Youden index (J) = **362 HU** (Sn= 57%, Sp=60%)

AUC= Area under the curve, PA= pulmonary artery, Sn= sensitivity, Sp= specificity