European Radiology (2022) 32:5831–5842
https://doi.org/10.1007/s00330-022-08645-2

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

How artificial intelligence improves radiological interpretation in suspected pulmonary embolism

Alexandre Ben Cheikh[1,2] · Guillaume Gorincour[1,3] · Hubert Nivet[1,4,5] · Julien May[1] · Mylene Seux[1] · Paul Calame[6,7] · Vivien Thomson[1,2] · Eric Delabrousse[6,7] · Amandine Crombé[1,8,9]

Received: 29 December 2021 / Revised: 29 December 2021 / Accepted: 4 February 2022 / Published online: 22 March 2022

Abstract

Objectives To evaluate and compare the diagnostic performances of a commercialized artificial intelligence (AI) algorithm for diagnosing pulmonary embolism (PE) on CT pulmonary angiogram (CTPA) with those of emergency radiologists in routine clinical practice.

Methods This was an IRB-approved retrospective multicentric study including patients with suspected PE from September to December 2019 (i.e., during a preliminary evaluation period of an approved AI algorithm). CTPA quality and conclusions by emergency radiologists were retrieved from radiological reports. The gold standard was a retrospective review of CTPA, radiological and clinical reports, AI outputs, and patient outcomes. Diagnostic performance metrics for AI and radiologists were assessed in the entire cohort and depending on CTPA quality.

Results Overall, 1202 patients were included (median age: 66.2 years). PE prevalence was 15.8% (190/1202). The AI algorithm detected 219 suspicious PEs, of which 176 were true PEs, including 19 true PEs missed by radiologists. In the cohort, the highest sensitivity and negative predictive values (NPVs) were obtained with AI (92.6% versus 90% and 98.6% versus 98.1%, respectively), while the highest specificity and positive predictive value (PPV) were found with radiologists (99.1% versus 95.8% and 95% versus 80.4%, respectively). Accuracy, specificity, and PPV were significantly higher for radiologists except in subcohorts with poor-to-average injection quality. Radiologists positively evaluated the AI algorithm to improve their diagnostic comfort (55/79 [69.6%]).

Conclusion Instead of replacing radiologists, AI for PE detection appears to be a safety net in emergency radiology practice due to high sensitivity and NPV, thereby increasing the self-confidence of radiologists.

Key Points

- *Both the AI algorithm and emergency radiologists showed excellent performance in diagnosing PE on CTPA (sensitivity and specificity ≥ 90%; accuracy ≥ 95%).*
- *The AI algorithm for PE detection can help increase the sensitivity and NPV of emergency radiologists in clinical practice, especially in cases of poor-to-moderate injection quality.*
- *Emergency radiologists recommended the use of AI for PE detection in satisfaction surveys to increase their confidence and comfort in their final diagnosis.*

Keywords Pulmonary embolism · Computed tomography angiography · Artificial intelligence · Sensitivity and specificity · Predictive value of tests

✉ Guillaume Gorincour
g.gorincour@imadis.fr

[1] IMADIS, 48 Rue Quivogne, 69002, Lyon, Bordeaux, Marseille, France

[2] Ramsay Générale de Santé, Clinique de la Sauvegarde, Lyon, France

[3] ELSAN, Clinique Bouchard, Marseille, France

[4] Centre hospitalier de Saintonge, Saintes, France

[5] Centre Aquitain d'Imagerie, Bordeaux, France

[6] Department of Radiology, Centre Hospitalier Universitaire de Besançon, Besançon, France

[7] Nanomedecine Laboratory, INSERM EA4662, University of Franche-Comte, Besançon, France

[8] University of Bordeaux, Bordeaux, France

[9] Department of Radiology, Pellegrin University Hospital, Bordeaux, France

Abbreviations

AI	Artificial intelligence
CI	Confidence interval
COVID-19	Coronavirus disease 2019
CTPA	CT pulmonary angiogram
DCNN	Deep convolutional neural network
EC	European conformity
FDA	Food and Drug Administration
NPV	Negative predictive value
PACS	Picture archive and communication system
PE	Pulmonary embolism
PPV	Positive predictive value

Introduction

Due to the ever-increasing numbers of hospital admissions and CT scans requested by emergency departments, reliable and rapid diagnosis and communication of results to referring physicians are becoming major challenges for radiologists. In the setting of suspected acute pulmonary embolism (PE), this is especially true, as early initiation of anticoagulation therapy is associated with better outcomes [1].

PE is the third most frequent acute cardiovascular syndrome, with an annual incidence rate ranging from 39 to 115 per 100,000 [2, 3]. During the coronavirus disease 2019 (COVID-19) pandemic, the PE incidence increased more than two-fold worldwide [4].

The diagnosis of PE is based on clinical presentation, D-dimer testing, and CT pulmonary angiogram (CTPA). Moreover, CTPA allows the evaluation of early patient risk by detecting right ventricle enlargement [2].

Artificial intelligence (AI) is becoming an essential tool to help radiologists interpret medical images [5]. Recent studies have shown encouraging results of deep convolutional neural networks (DCNNs) for detecting critical findings on CT scans, notably for intracranial hemorrhage or acute cerebral ischemia [6–11].

Regarding the diagnosis of PE on CTPA, AI is expected to play a key role in the emergency workflow. In our emergency teleradiological group, we have been working with a Food and Drug Administration (FDA) and European Conformity (EC)–approved AI algorithm based on DCNNs for several months (AIDOC). This algorithm provides automated detection of PE and flags the positive examinations for PE in the worklist. It has shown excellent preliminary results in the detection of PE [12], but the gold standard used did not allow us to draw conclusions on the respective accuracies of AI, radiologists alone, and in which situation AI could help radiologists. Herein, we hypothesized that the AI could complement emergency radiologists in their clinical practice.

Consequently, our primary objective was to evaluate and compare the diagnostic accuracies of radiologists alone (during their on-call duty) and AI alone in the detection of PE on CTPA based on a retrospective multicentric exploratory series preceding the routine implementation of AI in our teleradiological emergency workflow. The gold standard for AI and radiologist performances was the retrospective review of the CTPAs by one senior radiologist and one expert in AI with full access to clinical records, radiological reports, and AI predictions. Our secondary objectives were to determine the impact of AI implementation on the PE detection rate, satisfaction of radiologists, and interpretation duration.

Material and methods

Study design

This multicentric study was approved by the French national radiological review board (CRM-2103–146).

Three cohorts from three quarterly time periods were recruited from French emergency departments. The main cohort ("cohort-2019") consisted of all consecutive patients meeting the following inclusion criteria: adult patients with a clinical suspicion of PE between 2019–09–21 and 2019–12–24, request for CTPA by an emergency physician, available CTPA on picture archive and communication system (PACS), available interpretation by the radiologist, and available AI result. Since AI was not used in clinical practice over this period (except for internal evaluation), this cohort was used for our main objective, i.e., to compare the diagnostic performances of radiologists alone and AI alone.

The final AI implementation for clinical practice and webinars regarding the use of AI for pulmonary embolism detection were provided to all radiologists involved at Imadis on 2019–12–24 evening, which was the end of recruitment for cohort-2019.

Thus, two other cohorts were used for our secondary objectives, i.e., the impact of AI on PE detection rate and on interpretation duration. They consisted of all consecutive patients from 2018–10–01 to 2018–12–31 ("cohort-2018" or "pre-AI") and from 2020–06–01 to 2020–08–31 ("cohort-2020", between the 1st and 2nd waves of the COVID-19 pandemic in France, after full AI implementation) who met the following inclusion criteria: adult patients, request for a CTPA, available CTPA on PACS, and available report by the radiologist. Furthermore, the presence of abnormal findings on CT scan compatible or suspicious or typical of COVID-19 in patients from cohort-2020 was prospectively collected (i.e., equivalent of CO-RADS 3, 4, or 5) [13]. The difference between these two cohorts was that all radiologists had access to AI results before validating their reports in cohort-2020.

The number of recruiting emergency departments was 42 in 2018, 50 in 2019, and 73 in 2020.

Imaging protocol

CTPAs were performed using various 16-, 64-, or 80-detector row CT-scanners with a standardized protocol for all hospitals and bolus-tracking intravenous iodine contrast media at a rate of 3–4 mL/s (Omnipaque 350, GE Healthcare; Iomeron 400, Bracco Diagnostics; and Ultravist 370, Bayer Healthcare).

Radiological interpretation

The radiological interpretation protocol met the current French recommendations for teleradiology practice [14]. Requests including clinical data were received from the emergency departments by our dedicated interpretation centers (in Bordeaux, Lyon, and Marseille, France), and the indications and specific technical protocols for CTPA were systematically validated by a radiologist using dedicated software (ITIS, Deeplink Medical). Once the examination was completed, images were securely transferred through a virtual private network to the PACS (Carestream Health 12). Images were interpreted by one of the on-site radiologists in one of the interpretation centers. Our medical team consisted of 150 radiologists, including 104 senior radiologists (with ≥ 5 years of emergency imaging experience) and 46 junior radiologists (i.e., residents with 3–5 years of emergency imaging experience). Radiologists operated on-site rotations in groups of at least 5 per night, including one senior per site.

Radiological reports for any suspected PE are standardized. Radiologists systematically provided the following information: presence/absence of PE (defined as filling defects within the pulmonary arterial vasculature), presence/absence of respiratory artefacts limiting the interpretation, and quality of contrast media injection (good, average, or poor)

Furthermore, patient age and sex, imaging protocol, and radiological interpretation duration were retrieved.

AI implementation

The PE detection algorithm was provided by AIDOC Medical (version 1.0 in cohort-2019 and cohort-2020). The algorithm has been FDA- and EC-approved [15]. There was no financial support for this study. The algorithm principle is detailed in the study by Weikert et al [12, 16]. The practical implementation consisted of the transfer of deidentified examinations from a virtual machine installed at Imadis to the AIDOC data center for analysis. In the case of positive AI results, a color-encoded map was transferred into Imadis PACS (reidentification process) to enable visualization of the area suspected. Figure 1 shows examples of AIDOC outputs.

Consensual reading as the gold standard

Three months after the end of the inclusion period of cohort-2019, one senior radiologist (A.B.C., with 16 years of experience in emergency imaging) and one IT engineer specialized in AI (J.M.) reviewed all cohort-2019 cases. Based on AI output and on the radiologists' reports, they validated if a PE was actually present or not. If diagnosis remained doubtful, the reader requested the judgement of other senior radiologists and consulted medical records.

Moreover, the characteristics of the true PEs missed by AI and by radiologists were retrospectively reviewed.

Satisfaction survey

In September 2020, i.e., 9 months after AI introduction for routine diagnosis at Imadis, a satisfaction survey was sent to

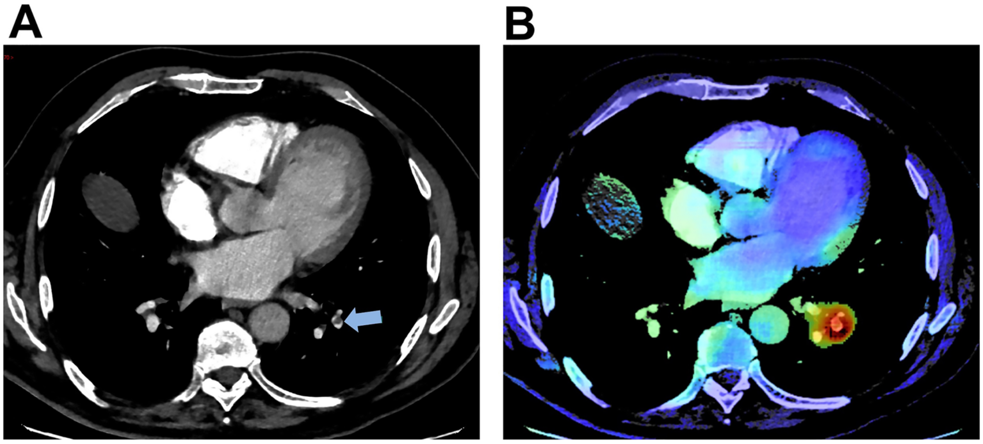

Fig. 1 Example of an output provided by the artificial intelligence (AI) algorithm (AIDOC Medical) to detect pulmonary embolism (PE). A 64-year-old man presented with spontaneous unilateral pain in the lower limb, increased with palpation, and unilateral edema. The revised simplified Geneva score was 3 and the D-dimer dosage was positive. **A** Contrast-enhanced CT pulmonary angiogram (CTPA) demonstrated a PE in the left lower limb (blue arrow). **B** On the same cross-section, the AI algorithm highlighted the same location of the suspected PE through a color-encoded map.

all radiologists with the following questions: "Does AI for PE detection improve diagnostic confidence?" and "Do you feel that AI for PE modifies your interpretation duration?" The answers corresponded to an ordered five-star scale from "*" (very negatively) to "*****" (very positively).

Statistical analyses

Statistical analyses were performed with R (v3.5.3). A p value < 0.05 was deemed significant.

Performances of AI and radiologists in cohort-2019 Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy (number of true positive and true negative observations divided by all the observations) with 95% confidence (95%CI) were estimated for the entire cohort.

Subgroup analysis These performance metrics were also estimated in the following subgroups of cohort-2019: (subcohort-1) examinations with average injection quality, (subcohort-2) examinations with poor injection quality, (subcohort-3) examinations with poor-to-moderate injection quality, (subcohort-4) examinations with respiratory artefacts limiting the interpretation, and (subcohort-5) examinations with respiratory artefacts and poor-to-moderate injection quality, as well as depending on the experience of radiologists (junior or senior).

Comparisons of AI and radiologists' performances Accuracies of AI and radiologists were compared with paired McNemar tests in the entire cohort-2019 and each subgroup.

Comparisons between pre-AI cohort-2018 and post-AI cohort 2019 The PE prevalence in the two cohorts was compared with the chi-squared test. After checking normality with the Shapiro–Wilk test, the interpretation durations between the two cohorts were compared with the unpaired Wilcoxon test. For this last test, examinations with protocols including additional exploration to the chest CTPA were removed to reduce bias, as well as patients from cohort-2020 with lung abnormalities compatible or strongly suspicious of COVID-19, as these two variables could bias and increase the interpretation time when positive.

Results

Study populations

Figure 2 shows the study flow chart. Overall, 1667 patients were included in cohort-2018 (median age 66.2 years, 961/1667 women [57.7%]), 1202 patients in cohort-2019 (median age 68.3 years, 689/1202 women [57.3%]), and 4454 in cohort-2020 (median age 68.6 years, 2236/4454 women [50.2%]) (Table 1 and Supplementary Data 1).

Regarding cohort-2019, 1558/2773 (56.2%) adult patients with CTPA were not post-processed with AI, and the diagnosis of PE remained equivocal for 13/1215 (1.1%) patients after consensual reading, leading to their exclusion (Fig. 2).

Thus, there were 190/1202 (15.8%) CTPAs with true PE (Table 1). The injection quality was rated as average and poor in 259/1202 (21.5%) and 67/1202 (5.6%) examinations. Significant respiratory artefacts were reported in 360/1202 (30%) examinations.

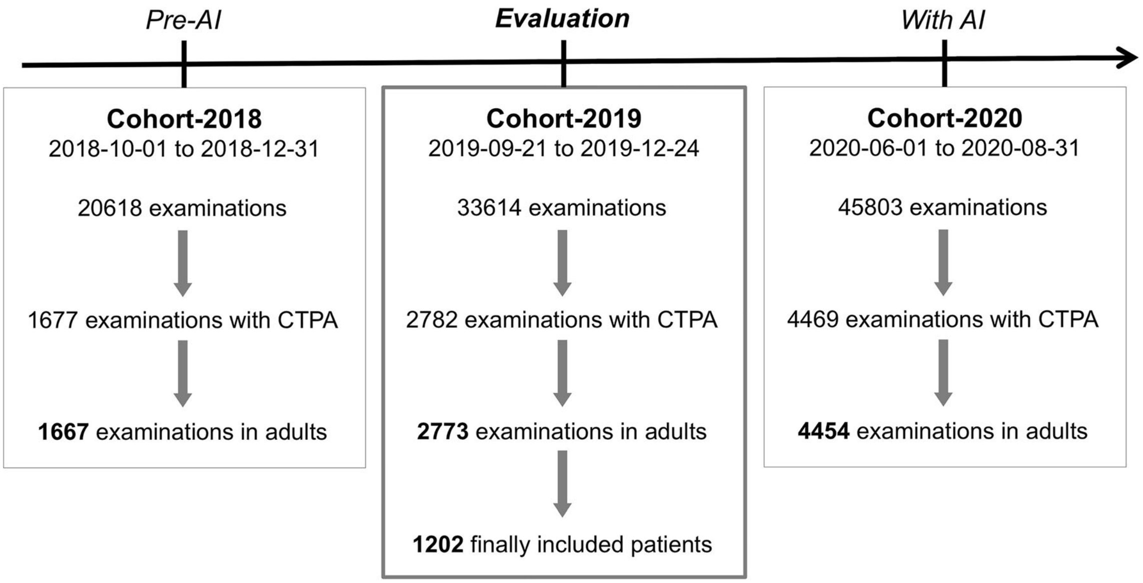
Pre-AI
Evaluation
With AI
Cohort-2018
2018-10-01 to 2018-12-31
20618 examinations
1677 examinations with CTPA
1667 examinations in adults
Cohort-2019
2019-09-21 to 2019-12-24
33614 examinations
2782 examinations with CTPA
2773 examinations in adults
1202 finally included patients
Cohort-2020
2020-06-01 to 2020-08-31
45803 examinations
4469 examinations with CTPA
4454 examinations in adults

Fig. 2 Flow chart of the study. Abbreviations: AI, artificial intelligence; CTPA, CT pulmonary angiogram

Table 1 Characteristics of cohort-2019

Characteristics	Patients
Age (years)	
Mean (sd)	65.4 ± 18.9
Median (range)	68.3 (18.1–101.9)
Sex	
Women	689/1202 (57.3%)
Men	513/1202 (42.7%)
Pulmonary embolism (: gold standard)	
Present	190/1202 (15.8%)
Absent	1012/1202 (84.2%)
Protocols	
CTPA	1100/1202 (91.5%)
CTPA + Abdomen-pelvic	78/1202 (6.5%)
Brain + CTPA	13/1202 (1.1%)
CTPA in pregnant women	8/1202 (0.7%)
Brain + CTPA + abdomen-pelvic	3/1202 (0.2%)
Respiratory artefacts limiting interpretation	
Yes	360/1202 (30%)
No	842/1202 (30%)
Quality of the injection	
Good	748/1202 (62.2%)
Average	259/1202 (21.5%)
Poor	67/1202 (5.6%)
No mention	128/1202 (10.6%)

Note. Results are number of patients with percentage in parentheses, except for age also given as mean ± standard deviation

Abbreviations: *CTPA* CT pulmonary angiogram, *sd* standard deviation

Table 2 summarizes the subcohorts derived from cohort-2019 with PE prevalence. The lowest PE prevalence was found in subcohort-2 (5/67 [7.5%]).

Performances of AI and radiologists to diagnose PE

The AI algorithm detected 219 suspicious PEs, of which 176 were true positives, 14 were false negatives, 43 were false positives, and 19 were true PEs missed by radiologists. The radiologists detected 180 suspicious PEs, of which 171 were true positives, 19 were false negatives, 9 were false positives, and 14 were true PEs missed by AI.

Table 3 shows the performance metrics of AI and radiologists in cohort-2019 and its subcohorts. Regarding the entire cohort, sensitivity and NPV were higher with AI than with radiologists, but the difference was not significant (sensitivity = 0.926 [95% CI: 0.879–0.959] with AI versus 0.900 [95% CI: 0.848–0.939] with radiologists, p = 0.3840; and NPV = 0.986 [95% CI: 0.977–0.991] with AI versus 0.981 [95% CI: 0.972–0.988] with radiologists, p = 0.4393, respectively).

Conversely, specificity and PPV were significantly higher with radiologists than with AI (specificity = 0.991 [95%CI: 0.983–0.996] versus 0.958 [95%CI: 0.943–0.969], p < 0.0001; and PPV = 0.950 [95%CI: 0.908–0.973] versus 0.804 [95%CI: 0.753–0.846], p < 0.0001, respectively).

Accuracy was significantly higher with radiologists than with AI (0.977 [95% CI: 0.967–0.984] versus 0.953 [95% CI: 0.939–0.964], p = 0.0024).

Regarding the subcohorts, similar trends were found, with sensitivity and NPV showing a tendency to be equal or higher with AI, without significant differences. Conversely, specificity and PPV were systematically higher with radiologists, except when the injection was of poor quality (subcohort-2; p = 0.3173 and 0.6015, respectively).

Accuracy was significantly higher with radiologists in subcohort-1, subcohort-4, and subcohort-5 (p = 0.0371, 0.003, and 0.0098, respectively, i.e., when the injection quality was average or when there were respiratory artefacts) but not in subcohort-2 and subcohort-3 (p = 1 and 0.0633, respectively, i.e., when the injection quality was poor or when the injection quality was poor-to-average).

Table 2 Summary of the subcohorts extracted from cohort-2019

Name	Properties	No. of patients	Positivity rate for PE
Cohort-2019	-	1202/1202 (100%)	190/1202 (15.8%)
Subcohort-1	CTPAs with average injection quality	259/1202 (21.5%)	38/259 (14.7%)
Subcohort-2	CTPAs with poor injection quality	67/1202 (5.6%)	5/67 (7.5%)
Subcohort-3	CTPAs with average-to-poor injection quality	326/1202 (27.1%)	43/326 (13.2%)
Subcohort-4	CTPAs with respiratory artefacts limiting the interpretation	360/1202 (30%)	44/360 (12.2%)
Subcohort-5	CTPAs with respiratory artefacts limiting the interpretation AND average-to-poor injection quality	168/1202 (14%)	18/168 (10.7%)

Note. Abbreviations: *CTPA* CT pulmonary angiogram, *no*. number; *PE* pulmonary embolism

Table 3 Performance statistics of artificial intelligence (AI) and radiologists in cohort-2019 and its subcohorts depending on factors limiting the radiological interpretation

Cohorts	Sensitivity (95%CI)	Specificity (95%CI)	PPV (95%CI)	NPV (95%CI)	Accuracy (95%CI)
Cohort-2019					
Radiologists	0.9 (0.848–0.939)	**0.991 (0.983–0.996)*****	**0.95 (0.908–0.973)*****	0.981 (0.972–0.988)	**0.977 (0.967–0.984)****
AI	**0.926 (0.879–0.959)**$^{\mathrm{ns}}$	0.958 (0.943–0.969)	0.804 (0.753–0.846)	**0.986 (0.977–0.991)**$^{\mathrm{ns}}$	0.953 (0.939–0.964)
Subcohort-1: Average injection quality					
Radiologists	0.895 (0.752–0.971)	**0.991 (0.968–0.999)****	**0.949 (0.823–0.987)*****	0.98 (0.952–0.992)	**0.977 (0.95–0.991)***
AI	**0.947 (0.823–0.994)**$^{\mathrm{ns}}$	0.932 (0.891–0.962)	0.724 (0.615–0.811)	**0.99 (0.961–0.997)**$^{\mathrm{ns}}$	0.934 (0.897–0.961)
Subcohort-2: Poor injection quality					
Radiologists	0.6 (0.147–0.947)	**0.984 (0.913–1)**$^{\mathrm{ns}}$	**0.875 (0.468–0.982)**$^{\mathrm{ns}}$	0.929 (0.817–0.975)	**0.955 (0.875–0.991)**
AI	**1 (0.478–1)**$^{\mathrm{ns}}$	0.952 (0.865–0.99)	0.697 (0.44–0.831)	**0.966 (0.866–0.988)**$^{\mathrm{ns}}$	**0.955 (0.875–0.991)**
Subcohort-3: Average-to-poor injection quality					
Radiologists	0.86 (0.721–0.947)	**0.989 (0.969–0.998)****	**0.938 (0.831–0.979)*****	0.974 (0.947–0.988)	**0.972 (0.948–0.987)**$^{\mathrm{ns}}$
AI	**0.953 (0.842–0.994)**$^{\mathrm{ns}}$	0.936 (0.901–0.962)	0.738 (0.642–0.816)	**0.991 (0.965–0.998)**$^{\mathrm{ns}}$	0.939 (0.907–0.962)
Subcohort-4: Limiting respiratory artifacts					
Radiologists	0.909 (0.783–0.975)	**0.994 (0.977–0.999)*****	**0.964 (0.871–0.991)*****	0.983 (0.958–0.993)	**0.983 (0.964–0.994)****
AI	**0.932 (0.813–0.986)**$^{\mathrm{ns}}$	0.937 (0.904–0.961)	0.734 (0.642–0.81)	**0.987 (0.961–0.995)**$^{\mathrm{ns}}$	0.936 (0.906–0.959)
Subcohort-5: Limiting respiratory artifacts AND average-to-poor injection quality					
Radiologists	**0.944 (0.727–0.999)**	**0.993 (0.963–1)****	**0.964 (0.79–0.995)*****	**0.99 (0.934–0.998)**$^{\mathrm{ns}}$	**0.988 (0.958–0.999)***
AI	**0.944 (0.727–0.999)**	0.92 (0.864–0.958)	0.689 (0.56–0.794)	0.989 (0.929–0.998)	0.923 (0.871–0.958)

Note. Abbreviations: *95%CI* 95% confidence interval, *AI* artificial intelligence, *NPV* negative predictive value, *PPV* positive predictive value, *ns* not significant; *: $p < 0.05$; **: $p < 0.005$; ***: $p < 0.001$. For each metrics and each subcohort; the highest value between AI and radiologists is indicated in boldface

Figure 3 represents these performance statistics for the entire cohort-2019 and subcohort-2 (poor-quality injection).

Junior and senior radiologists interpreted 301/1202 (25%) and 901/1202 (75%) CTPAs, respectively. The accuracies of seniors and juniors were similar (0.977 [95%CI: 0.965–0.986] and 0.977 [95%CI: 0.953–0.991], respectively) and higher than the accuracy of AI in the two corresponding subgroups (0.953 [95%CI: 0.938–0.966] and 0.950 [95%CI: 0.919–0.972], respectively) (Supplementary Data 2).

Table 4 displays the absolute number and proportions of discrepancies between AI and radiologists, AI and gold standard, and radiologists and gold standard for each cohort. Conversely, AI falsely indicated PE in 34 patients, while the radiologists correctly indicated the absence of PE. The highest rate of discordance between radiologists and AI was found in subcohort-2 (9% [6/67]), which was also the subcohort in which radiologists had the highest discordance with the gold standard (4.5% [3/67]).

Table 5 summarizes the characteristics of true PEs missed by radiologists and not by AI and missed by AI and not by radiologists. Overall, PEs missed by radiologists were interpreted within a shorter time than PEs missed by AI (p = 0.0390), more frequently present along with other confusing thoracic diseases (p = 0.0265), and characterized by shortest clot length (p = 0.0186).

Figure 4 illustrates PE diagnosed by AI and not radiologists, and PE diagnosed by a radiologist and not AI.

Clinical use of AI by radiologists

Figure 5 A, B shows the results of the satisfaction survey. Responses were provided by 79/150 (52.7%) Imadis radiologists. Regarding improvement in diagnostic confidence, 57/79 (72.2%) radiologists found that the availability of AI was positive or strongly positive. Regarding the interpretation duration, most radiologists were neutral (41/79, 51.9%), while 27/79 (34.2%) found that using AI increased it.

The prevalence of PE was 16.3% (256/1667) in cohort-2018 versus 12.7% (566/4454) in cohort-2020, which was significantly different (p = 0.0077) (Supplemental Data 1). The average interpretation duration for CTPA alone was 14.55 ± 9.08 min in patients from cohort-2018 (n = 1459) versus 15.6 ± 9.77 min in patients from cohort-2020 (without findings compatible or typical of COVID-19, n = 3311), which was significantly different (p < 0.0001) (Fig. 5C).

Discussion

Herein, we compared the diagnostic performance of a validated AI algorithm for PE detection on CTPA with those of

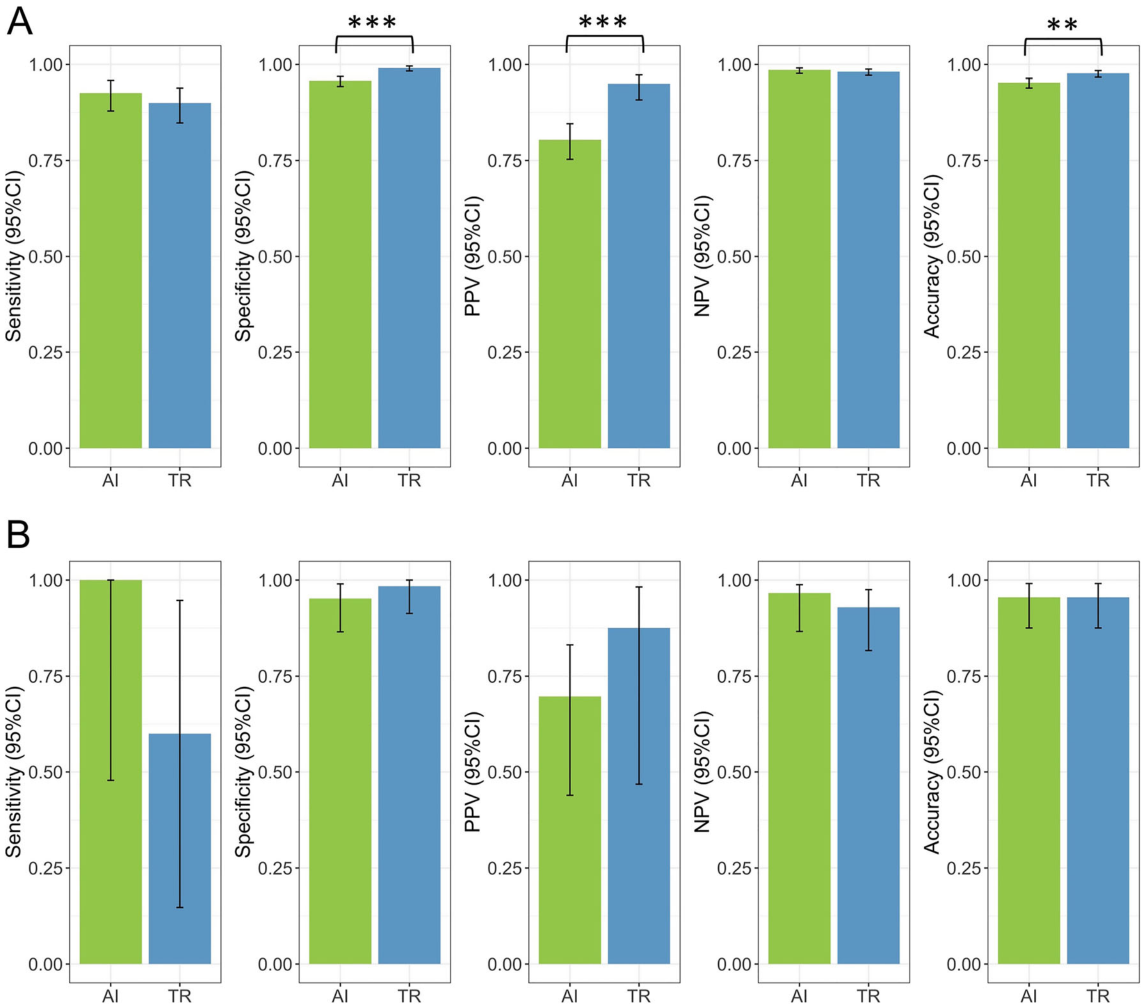

Figure 3 Performances of artificial intelligence (AI) and teleradiologists (TR) to diagnose pulmonary embolism on CT pulmonary angiogram (CTPA) in a multicentric emergency cohort. **A** Patients from the entire cohort-2019. **B** Patients from the subcohort 2 with poor quality injection. Abbreviations: 95%CI, 95% confidence interval; NPV, negative predictive value; PPV, positive predictive value. *: $p < 0.05$, **: $p < 0.005$; ***: $p < 0.001$

Table 4 Discordances between radiologists, artificial intelligence (AI), and gold-standard in cohort-2019 and its subcohorts depending on factors limiting the radiological interpretation.

Cohorts	Discordance between AI and radiologists	Discordance between AI and gold-standard	Discordance between radiologists and gold-standard	No. of TPs captured by AI and not by radiologists	No. of TPs captured by radiologists and not by AI
Cohort-2019	85/1202 (7.1%)	57/1202 (4.7%)	28/1202 (2.3%)	19/190 (10%)	14/190 (7.4%)
Subcohort 1 (average injection quality)	23/259 (8.9%)	17/259 (6.6%)	6/259 (2.3%)	4/38 (10.5%)	2/38 (5.3%)
Subcohort 2 (poor injection quality)	6/67 (9%)	3/67 (4.5%)	3/67 (4.5%)	2/5 (40%)	0/5 (0%)
Subcohort 3 (average or poor injection quality)	29/326 (8.9%)	20/326 (6.1%)	9/326 (2.8%)	6/43 (14%)	2/43 (4.7%)
Subcohort 4 (respiratory artifacts)	29/360 (8.1%)	23/360 (6.4%)	6/360 (1.7%)	4/44 (9.1%)	3/44 (6.8%)
Subcohort 5 (respiratory artifacts AND average or poor injection quality)	15/168 (8.9%)	13/168 (7.7%)	2/168 (1.2%)	1/18 (5.6%)	1/18 (5.6%)

Note. Abbreviations: *AI* artificial intelligence, *FP* false positive, *TP* true positive

Table 5 Review of the true pulmonary embolism (PE) missed by artificial intelligence (AI) and radiologists

Characteristics	PE missed by AI ($n = 14$)	PE missed by radiologists ($n = 19$)	p value
Radiologists' experience			
Junior	4/14 (28.6%)	3/19 (15.8%)	0.6477
Senior	10/14 (71.4%)	16/19 (84.2%)	
Interpretation duration (min)[1]	23 ± 13	14.4 ± 7	**0.0390***
Patients' sex			
Women	8/14 (57.1%)	12/19 (63.2%)	1
Men	6/14 (42.9%)	7/19 (36.8%)	
Patients' age (years)[1]	65.9 ± 18.7	75.3 ± 14.2	0.1353
Respiratory artifacts			
Absent	7/10 (70%)	13/17 (76.5%)	1
Present	3/10 (30%)	4/17 (23.5%)	
Density in pulmonary trunk (HU)[1]	432.3 ± 188	440.6 ± 156.1	0.8555
Presence of other confusing abnormal findings			
No	10/14 (71.4%)	5/19 (26.3%)	**0.0265***
Yes	4/14 (28.6%)	14/19 (73.7%)	
Details regarding the abnormal findings			
Pleural effusion	3/14 (21.4%)	3/19 (15.8%)	1
Infectious disease	4/14 (28.6%)	7/19 (36.8%)	0.9009
Cardiac decompensation	0/14 (0%)	2/19 (10.5%)	0.607
Neoplasia	1/14 (7.1%)	2/19 (10.5%)	1
Atelectasia	2/14 (14.3%)	4/19 (21.1%)	0.9669
Emphysema	0/14 (0%)	3/19 (15.8%)	0.3438
Major dorsal kyphosis	1/14 (7.1%)	2/19 (10.5%)	1
Other	1/14 (7.1%)	1/19 (5.3%)	1
PE laterality			
Right	6/14 (42.9%)	9/19 (47.4%)	0.3701
Left	3/14 (21.4%)	7/19 (36.8%)	
Bilateral	5/14 (35.7%)	3/19 (15.8%)	
Number of lung lobes with PE[1]	1.6 ± 1.1	1.2 ± 0.5	0.2144
Number of clot			
Multiple	7/14 (50%)	5/19 (26.3%)	0.3022
Single	7/14 (50%)	14/19 (73.7%)	
Location of the most proximal clot			
Proximal	3/14 (21.4%)	0/19 (0%)	0.1802
Lobar	2/14 (14.3%)	3/19 (15.8%)	
Segmental	7/14 (50%)	14/19 (73.7%)	
Sub-segmental	2/14 (14.3%)	2/19 (10.5%)	
Positioning of the clot in the vessel			
Central	13/14 (92.9%)	15/19 (78.9%)	0.5417
Marginal	1/14 (7.1%)	4/19 (21.1%)	
Occlusive clot			
No	5/14 (35.7%)	12/19 (63.2%)	0.2276
Yes	9/14 (64.3%)	7/19 (36.8%)	
Length of the longest clot (mm)[1]	21.4 +/- 14.1	10.7 +/- 6.9	**0.0186***
Dilated pulmonary artery			
No	8/14 (57.1%)	15/19 (78.9%)	0.3351
Yes	6/14 (42.9%)	4/19 (21.1%)	

Note. Data are number of patients with percentage in parentheses except for ([1]), where data are mean ± standard deviation. The tests are chi-square or Fisher test for categorical characteristics and unpaired Wilcoxon test for numerical characteristics

Other abbreviations: *HU*, Hounsfield unit

*: $p < 0.05$

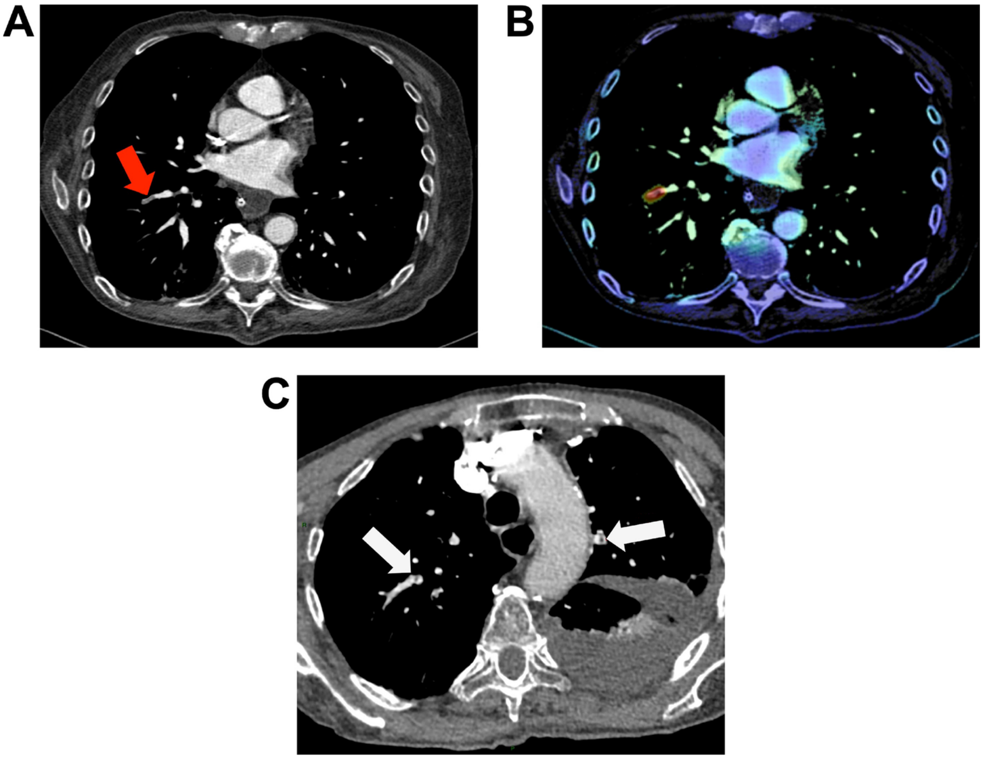

Figure 4 Clinical examples. A 71-year-old patient with a medical history of cancer and recent surgery presented with heart rate > 95 beats per minute and a borderline saturation and underwent a contrast-enhanced CT pulmonary angiogram (CTPA) (**A**), which showed a segmental, sub-acute, pulmonary embolism (PE) in the right low limb, which was missed by the emergency radiologist during his on-call duty (red arrow). **B** On the same cross-section, the PE was correctly identified by the artificial intelligence (AI) algorithm (AIDOC Medical). Example of pulmonary embolism (PE) correctly diagnosed by the emergency radiologist and not by the artificial intelligence (AI) algorithm. Opposite example: An 85-year-old patient with a medical history of PE and a recent surgery presented with a heart rate between 75 and 94 beats per minute and acute dyspnea and underwent CTPA (**C**). Two segmental PEs were correctly diagnosed by the emergency radiologist but missed by the AI algorithm (white arrows)

radiologists in a clinical setting. In the entire cohort, sensitivity and NPV were higher with AI than with radiologists, although the differences were not significant. Conversely, specificity, PPV, and accuracy were significantly higher with radiologists than with AI. However, our results were more contrasted for poor-quality examinations and for the appreciation of AI tools

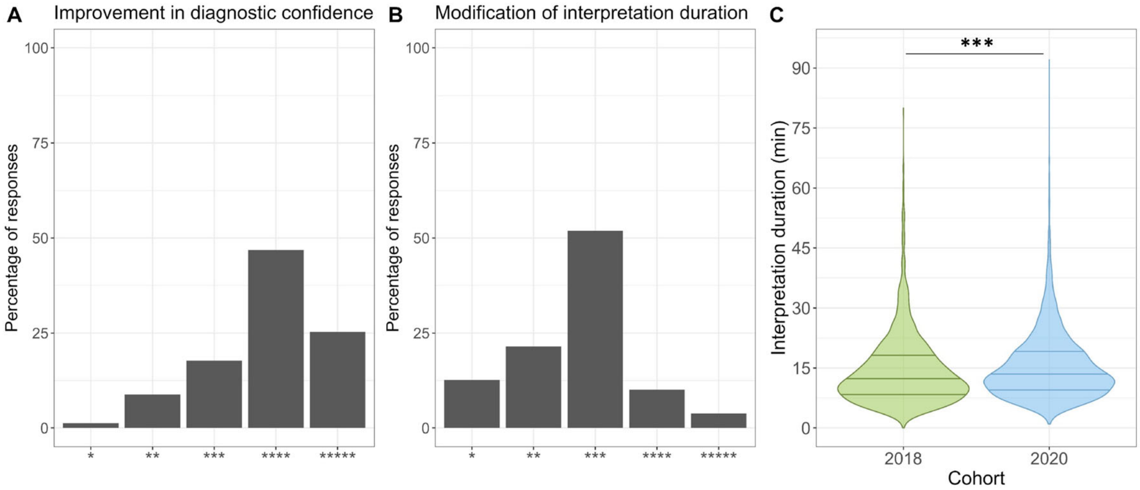

Figure 5 Use of artificial intelligence (AI) by radiologists for emergency clinical routine at Imadis. Qualitative assessment: results of the satisfaction survey sent 9 months after implementing AI in clinical workflow (**A**, **B**). Quantitative assessment (**C**): comparison of interpretation duration for a single CT pulmonary angiogram in 2018 (without AI) and 2020 (with AI) (lines inside the violin plots correspond to 1st quartile, median, and 3rd quartile). ***: $p < 0.001$

by radiologists. Indeed, radiologists stressed the importance of AI to strengthen their conclusions, especially to confirm negative findings, or to ensure the absence of distal PE in poor-quality examinations.

Previous computer-assisted detection solutions for automatic detection of PE were hampered by long calculation times, rather low sensitivities [16–20], and low specificities [21–30], leading to increased workloads for radiologists [31]. These disappointing performances were likely due to the datasets used to train the AI algorithms, which were often imbalanced towards high PE prevalence or sometimes even included only positive examinations [16, 18, 19, 22, 26, 28, 31]. Thus, the clinical implementation of algorithms with acceptable computational time, easily includable in emergency workflows, represents a major breakthrough.

In the study by Weikert et al performed on a comparable number of patients, the same AI algorithm correctly identified 215 of 232 PEs, providing 92.7% sensitivity, which is comparable to ours (92.6%). Furthermore, the AI algorithm reported 1178 of 1233 negative CTPAs, providing 95.5% specificity, again comparable to ours (95.8%). In their study, the gold standard corresponded to written radiological reports approved by at least two radiologists, at least one of them being board-certified. The visual review of AI outputs was performed by a 3rd year radiology resident. Indeterminate findings were read by a board-certified radiologist with 4 years of professional experience. Negative examinations according to both radiological reports and AI algorithms were considered true negatives and not checked visually. However, the authors did not compare the performances of the AI algorithm with those of the radiologists during their clinical activity (herein, emergency on-call period), did not review the failures of radiologists, and did not investigate the changes in interpretation time and PE prevalence following AI implementation, which has been performed in our work. Thus, regarding our first objective, our results suggest that the AI algorithm did not demonstrate significantly better diagnostic performance, regardless of the metrics used. However, our secondary objectives enable us to propose a more contrasting analysis.

Indeed, in the entire cohort-2019, AIDOC captured 19 PEs that were not diagnosed by radiologists in 19 distinct patients. In other words, the AI algorithm could correct a misdiagnosed PE approximately every 63 CTPAs (≈1202/19). This estimation must be considered in parallel with the high number of CTPAs required by emergency physicians (≈18,000 CTPAs in 2020 in our group—so approximately 285 [≈18000/1202 × 19] true PEs detected by AI but initially misdiagnosed by radiologists in 2020) and with human and financial consequences of missed PEs [32]. Indeed, mortality and recurrence rates for untreated or missed PE range between 5 and 30%.

Furthermore, our subgroup analysis highlighted the similar accuracy for junior and senior radiologists, as well as the interest of AI for poor-quality examinations, especially regarding the injection. Indeed, although the number of patients in this subgroup was small (subcohort-2, $n = 67$), we observed higher sensitivity, NPV, and similar accuracy with AI. Interestingly, the highest number of discrepancies between radiologists and the gold standard was found in subcohort-2, in agreement with previous studies [19, 33, 34]. Indeed, Miller reported that rates of false-positive and indeterminate examinations increased with diminishing CTPA quality. Thus, we believe that supporting radiological diagnosis in these particular settings (where radiologists can lack confidence and be less efficient) could be a main advantage of AI. In addition, acquisitions with low image quality remain practically challenging. In these situations, the additional uses of CT scan improvement via AI [35], as well as spectral imaging [36], could be valuable. Interestingly, the retrospective review of missed PEs suggested that radiologists' failures were also partly due to distraction by concomitant confusing thoracic diseases.

Regarding secondary objectives, thanks to the comparison of cohort-2018 and cohort-2020, we found that the average interpretation time for CTPA alone increased by 1 min 3 s (7.2+%) since AI clinical implementation (after controlling potential confusion bias such as multiple acquisitions and COVID-19). In contrast, Duron recently found that using AI for detecting adult appendicular skeletal fractures decreased the interpretation time by 15% [34]. The nature of medical images can explain these findings (2D radiographs versus 3D CT scans). Moreover, a teleradiology workflow leads to the transfer of a large volume of data. Bandwidth limitations and the level of security required can also explain the delay.

Additionally, the satisfaction survey (performed several months after AI clinical implementation) highlighted the positive reception by radiologists who emphasized that AI has increased their diagnostic confidence and comfort when interpreting CTPAs. The higher sensitivity and NPV of AI are probably the underlying reasons for this additional confidence, especially in late night, while correcting falsely positive predictions made by AI is generally rapidly achieved. Another secondary interest of implementing FDA- and CE-approved AI algorithms in clinical practice could be to protect radiologists from medical malpractice litigations. Thus, beyond the raw comparisons of diagnostic performances, the overall balance between the advantages and disadvantages appears to favor the use of AI in clinical practice to complement (and possibly to augment) radiologists' interpretations.

Our work has limitations. First, it was retrospective. Second, we did not compare the diagnostic performances of AI alone and radiologists alone with those of radiologists with full access to the AI output during their practice (i.e., "augmented" radiologists). Only a prospective comparative study could achieve this and enable to draw a conclusion. Third, we did not investigate other features that could influence the performances of radiologists but not AI, such as interpretation hours (especially during late nights) or high workload periods. The type of scanner, its acquisition, and post-processing parameters could have also influenced the performances of AI and radiologists, but the number of patients per center was too small for this sub-analysis. Moreover, some features could alter the performances of both AI and radiologists, such as the location of PE in the lung, its distality, or its location within the vascular lumen [12]. Fourth, we did not observe an increase in the diagnosis of PE following AI clinical implementation, as shown in comparisons between cohort-2018 and cohort-2020. However, cohort-2020 might have been biased by the COVID-19 pandemic, which resulted in more requests for chest CT, overall [37], and specifically for suspected PE [38]—though these rises were also due to the increase in partner centers of IMADIS. The injection quality was rated as poor in only 5.6% of examinations, whereas respiratory artefacts were reported in 30%. This discrepancy can be explained by the high level of protocol homogenization needed in teleradiology settings, reaching its limits in the absence of on-site radiologists. It also reflects the difficulties for radiographers to obtain technically good examinations in a real-life multicentric cohort of symptomatic and dyspneic patients from the emergencies during the on-call period.

In conclusion, this study confirms the high diagnostic performances of AI algorithms relying on DCNN to diagnose PE on CTPA in a large multicentric retrospective emergency series. It also underscores where and how AI algorithms could better support (or "augment") radiologists, i.e., for poor-quality examinations and by increasing their diagnostic confidence through the high sensitivity and high NPV of AI. Thus, our work provides more scientific ground for the concept of "AI-augmented" radiologists instead of supporting the theory of radiologists' replacement by AI.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-022-08645-2.

Acknowledgements We would like to sincerely thank all our partner centers and the support team at IMADIS and especially Lucie Vauthier.

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

Declarations

Guarantor The scientific guarantor of this publication is Dr Guillaume Gorincour.

Conflict of interest The authors of this manuscript declare relationships with the following companies: ABC, GG, VT have shares in DeepLink Medical. ABC and VT have shares in Gleamer.

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

Statistics and biometry One of the authors has significant statistical expertise: AC.

Informed consent Written informed consent was not required for this study because of the retrospective nature of this study, patients were informed about the reuse of their anonymized data.

Ethical approval Institutional Review Board approval was obtained (CERIM: CRM-2103-146).

Methodology
- retrospective
- observational
- multicenter study

References

1. Smith SB, Geske JB, Maguire JM et al (2010) Early anticoagulation is associated with reduced mortality for acute pulmonary embolism. Chest 137:1382–1390
2. Konstantinides SV, Meyer G, Becattini C et al (2019) 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). Eur Heart J 41:543–603
3. Essien EO, Rali P, Mathai SC (2019) Pulmonary Embolism. Med Clin North Am 103:549–564
4. Suhail Akhter M, Hamali HA, Mobarki AA et al (2021) Clinical medicine SARS-CoV-2 infection: modulator of pulmonary embolism paradigm. J Clin Med 10:1064
5. Barragán-Montero A, Javaid U, Valdés G et al (2021) Artificial intelligence and machine learning for medical imaging: a technology review. Phys Medica 83:242–256
6. Prevedello LM, Little KJ, Qian S, White RD (2017) Artificial intelligence in imaging 1. Radiology 285:923–931
7. Lee JY, Kim JS, Kim TY, Kim YS (2020) Detection and classification of intracranial haemorrhage on CT images using a novel deep learning algorithm. Sci Rep 10:20546 1–7. https://doi.org/10.1038/s41598-020-77441-z
8. Nagel S, Sinha D, Day D et al (2016) e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. Int J Stroke 0:1–8. https://doi.org/10.1177/1747493016681020
9. Shi Z, Miao C, Schoepf UJ et al (2020) A clinically applicable deep-learning model for detecting intracranial aneurysm in computed tomography angiography images. Nat Commun. 11:6090. https://doi.org/10.1038/s41467-020-19527-w
10. Winkel DJ, Heye T, Weikert TJ et al (2019) Evaluation of an AI-based detection software for acute findings in abdominal computed tomography scans: toward an automated work list prioritization of routine CT examinations. Invest Radiol 54:55–59
11. Gorincour G, Monneuse O, Ben CA et al (2021) Management of abdominal emergencies in adults using telemedicine and artificial intelligence. J Visc Surg 158:S26–S31. https://doi.org/10.1016/j.jviscsurg.2021.01.008
12. Weikert T, Winkel DJ, Bremerich J et al (2020) Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. Eur Radiol. 30(12):6545–6553. https://doi.org/10.1007/s00330-020-06998-0
13. Prokop M, van Everdingen W, van Rees VT et al (2020) CO-RADS: a categorical CT assessment scheme for patients suspected of having COVID-19-definition and evaluation. Radiology 296(2):E97–E104

14. Société Française de Radiologie (2018) Qualité et sécurité des actes de téléimagerie – Guide de bonnes pratiques. http://www.sfrnet.org/sfr/professionnels/2-infos-professionnelles/05-teleradiologie/index.phtml. Accessed 4 Apr 2020
15. AIDOC (2020) Pulmonary embolism guidelines and the intersection with AI. https://www.aidoc.com/blog/pulmonary-embolism-guidelines-and-the-intersection-with-ai/. Accessed 1 June 2020
16. Maizlin ZV, Vos PM, Godoy MB, Cooperberg PL (2007) Computer-aided Detection of Pulmonary Embolism on CT Angiography Initial Experience. J Thorac Imaging 22:324–329. https://doi.org/10.1097/RTI.0b013e31815b89ca
17. Bouma H, Sonnemans JJ, Vilanova A, Gerritsen FA (2009) Automatic detection of pulmonary embolism in CTA images. IEEE Trans Med Imaging 28:1223–1230. https://doi.org/10.1109/TMI.2009.2013618
18. Brown JB, Gestring ML, Leeper CM et al (2017) The value of the injury severity score in pediatric trauma: time for a new definition of severe injury? J Trauma Acute Care Surg. https://doi.org/10.1097/TA.0000000000001440
19. Lee CW, Seo JB, Song J et al (2011) Evaluation of computer-aided detection and dual energy software in detection of peripheral pulmonary embolism on dual-energy pulmonary CT angiography. Eur Radiol:54–62
20. Lee G, Lee HY, Park H et al (2016) Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management : state of the art. Eur J Radiol. https://doi.org/10.1016/j.ejrad.2016.09.005
21. Masutani Y, Macmahon H, Doi K (2002) Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis. IEEE Trans Med Imaging 21:1517–1523. https://doi.org/10.1109/TMI.2002.806586
22. Pichon E, Novak CL, Kiraly AP, Naidich DP (2004) A novel method for pulmonary emboli visualization from high-resolution CT images. Proc. SPIE 5367, Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display, (5 May 2004); Robert L. Galloway Jr., Editor(s). https://doi.org/10.1117/12.532892
23. Liang J, Bi J (2007) Computer aided detection of pulmonary embolism with tobogganing and multiple instance classification in CT pulmonary angiography. Inf Process Med Imaging 20:630–641
24. Digumarthy SR, Kagay CR, Legasto AC, et al (2006) Computer-aided detection (CAD) of acute pulmonary emboli: evaluation in patients without significant pulmonary disease. In: Radiological Society of North America 2006 Scientific Assembly and Annual Meeting. Chicago IL
25. Schoepf UJ, Schneider AC, Das M et al (2007) Pulmonary embolism : computer-aided detection at multidetector row spiral computed tomography. J Thorac Imaging 22:319–323
26. Wittenberg R, Peters JF, Sonnemans JJ et al (2010) Computer-assisted detection of pulmonary embolism : evaluation of pulmonary CT angiograms performed in an on-call setting. Eur Radiol 20: 801–806
27. Blackmon KN, Florin C, Bogoni L et al (2011) Computer-aided detection of pulmonary embolism at CT pulmonary angiography : can it improve performance of inexperienced readers ? Eur Radiol 21:1214–1223
28. Wittenberg R, Berger FH, Peters JF et al (2012) Acute pulmonary embolism : effect of a computer-assisted detection prototype on purpose: methods: results. Radiology 262. https://doi.org/10.1148/radiol.11110372
29. Lahiji K, Kligerman S, Jeudy J, White C (2014) Improved Accuracy of pulmonary embolism computer-aided detection using iterative reconstruction compared with filtered back projection. AJR Am J Roentgenol:763–771
30. Wittenberg R, Peters JF, Van Den BIAH et al (2013) Computed tomography pulmonary angiography in acute pulmonary embolism -the effect of a computer-assisted detection prototype used as a concurrent reader. J Thorac Imaging 28:315–321
31. Bhargavan M, Kaye AH, Forman HP, Sunshine JH (2009) Workload of radiologists in United States in 2006–2007 and trends since 1991–1992. Radiology 252. https://doi.org/10.1148/radiol.2522081895
32. Calder KK, Herbert M, Henderson SO (2005) The mortality of untreated pulmonary embolism in emergency department patients. Ann Emerg Med 45:302–310
33. Das M, Mühlenbruch G, Helm A et al (2008) Computer-aided detection of pulmonary embolism : influence on radiologists ' detection performance with respect to vessel segments. Eur Radiol: 1350–1355
34. Miller WTJ, Arinari LA, Barbosa EJ et al (2015) Small pulmonary artery defects are not reliable indicators of pulmonary embolism. Ann Am Thorac Soc. https://doi.org/10.1513/AnnalsATS.201502-105OC
35. Leithner D, Gruber-Rouh T, Beeres M et al (2018) 90-kVp low-tube-voltage CT pulmonary angiography in combination with advanced modeled iterative reconstruction algorithm: Effects on radiation dose, image quality and diagnostic accuracy for the detection of pulmonary embolism. Br J Radiol 91. https://doi.org/10.1259/bjr.20180269
36. Lysdahlgaard S, Hess S, Gerke O, Weber Kusk M (2020) A systematic literature review and meta-analysis of spectral CT compared to scintigraphy in the diagnosis of acute and chronic pulmonary embolisms. Eur Radiol 30:3624–3633. https://doi.org/10.1007/s00330-020-06735-7
37. Crombé A, Lecomte J-C, Banaste N et al (2021) Emergency teleradiological activity is an epidemiological estimator and predictor of the COVID-10 pandemic in mainland France. Insights Imaging 12(1):103
38. Grillet F, Busse-Coté A, Calame P et al (2020) COVID-19 pneumonia: microvascular disease revealed on pulmonary dual-energy computed tomography angiography. Quant Imaging Med Surg 10: 1852–1862