



Value assessment of augmentative artificial intelligence for assessment of pulmonary emboli on CT – a meta-analysis comprising 15,963 CT scans

Igor M. Kitanovski¹ · Alec Buetow² · Steven C. Schoettler-Woll³  · Abdul M. Zafar⁴

Received: 4 March 2025 / Accepted: 7 April 2025

© The Author(s), under exclusive licence to American Society of Emergency Radiology (ASER) 2025

Abstract

Purpose Artificial Intelligence (AI) algorithms in radiology are currently deployed as tools to augment radiologists rather than autonomous readers. An augmentative tool should improve performance above and beyond the baseline performance achieved by the user alone. We conducted a meta-analysis to elucidate the added value of augmentative AI to radiologists for detecting Pulmonary Embolism (PE) on CT scan.

Methods Using PRISMA guidelines, studies in which both AI and Human Interpreter (HI) assessed CT scans for pulmonary emboli were selected. Data extracted from these studies were used to compare diagnostic performance of AI and HI with an emphasis on the performance of AI above and beyond that of HI.

Results Both HI and AI performed similarly with no statistically significant difference in the pooled estimates of sensitivity, specificity, PPV, NPV and accuracy. Subsequent analysis focusing on the differences between performance of AI and HI within each study, followed by pooled estimate, also did not demonstrate any significant difference ($p < 0.05$).

Conclusions In a meta-analysis of nearly sixteen thousand CTs, AI and HI had similar performance for detection of pulmonary emboli. On one hand, this buttresses AI's use for triaging and for second reads. On the other hand, the outcomes may or may not be different when AI is added-on. The findings of this meta-analysis can be used to re-examine the use-scenarios of AI and to re-calibrate its value proposition.

Keywords Augmented intelligence · Radiology · Pulmonary embolism · CT scan

Introduction

The worldwide shortage of adequately trained radiologists, years-long training required for practicing radiology (e.g. minimum of 13 years of post-secondary training in USA), and the potential for a reader free from the human limitations (oculomotor fatigue, sleepiness, distractions etc.)

make Artificial Intelligence (AI) in radiology a promising avenue which has garnered great interest and investments over the past decade. Currently, close to 700 AI imaging algorithms have been cleared by the FDA [1]. A growing body of literature demonstrates similar performance of AI and human interpreters (HI) for multiple imaging findings such as pulmonary emboli, intracranial bleeds, and spinal fractures [2].

Under the current regulatory framework, AI algorithms in radiology are deployed as augmentative tools rather than autonomous readers. The U.S. Food and Drug Administration (FDA) guidelines stipulate that all FDA-cleared AI devices for diagnostic radiology are intended to be used as adjuncts to radiologists, not as replacements. This approach ensures that responsibility and accountability remain with human operators [3, 4]. European and North American radiology societies also corroborate the use of AI in an augmentative capacity with physician oversight in order to ensure

✉ Steven C. Schoettler-Woll
steven.c.schoettler.woll@hitchcock.org

¹ Dartmouth Geisel School of Medicine, Hanover, NH, USA

² Department of Epidemiology, Dartmouth Geisel School of Medicine, Hanover, NH, USA

³ Dartmouth-Hitchcock Medical Center, Radiology Department, Lebanon, NH, USA

⁴ Radiology Department, Dartmouth Geisel School of Medicine, Hanover, NH, USA

that patient care decisions are made with a comprehensive understanding of individual contexts [5, 6].

In the current practice, AI flags findings for review by an HI and appears very similar to a high-functioning computer aided detection (CAD) software. AI is often limited to one scenario per algorithm e.g. one algorithm assesses pulmonary emboli (PE), a second algorithm assesses pneumonia and a third algorithm assesses rib fractures on the same study. In certain use scenarios, AI elevates the priority of an imaging study so it can be reviewed earlier by HI, and by extension, moves the other patients' studies to lower priority. Augmentative AI (Aug-AI) assumes no liability for the accuracy of its findings or the outcomes of its actions. This responsibility remains with the HI.

A tool, such as Aug-AI, should have a net positive value above and beyond that afforded by the human user alone to justify the time, effort and resources required for its development and deployment. When Aug-AI correctly classifies a finding initially misclassified by HI, Aug-AI adds value and potentially improves outcomes. On the contrary, when Aug-AI misclassifies a study correctly classified by the HI, AI subtracts value by consuming HI's time and effort to resolve the discrepancy. When both Aug-AI and HI make the same call (right or wrong), the net effect is neutral. Most of the literature on diagnostic AI has focused on a direct comparison between HI and AI which would be more suited to an autonomous AI. We performed a meta-analysis of the recent literature to assess the diagnostic performance of AI for detecting pulmonary emboli on CT with a focus on the difference between the performance of AI and HI reading the same studies.

Methods

This meta-analysis followed Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [7].

Search strategy

Four databases (PubMed, OVID, SCOPUS, and WebOf-Science) were searched between September 2024 and October 2024 using the following search terms: "Artificial intelligence and pulmonary emboli", "AI in the detection of pulmonary emboli", "artificial intelligence AND pulmonary embolism AND improvement", and "AI AND pulmonary embolism AND human interpreter", "Artificial intelligence AND pulmonary embolism AND radiologist", "radiology AND pulmonary embolism AND artificial intelligence", "AI pulmonary embolism detection", and "Artificial intelligence AND pulmonary embolism AND comparison". Since

AI is a rapidly evolving field, the search was limited to studies published in the last 3 years i.e. September 1st, 2021, and August 31st, 2024. The search was limited to English language manuscripts which had abstracts.

Screening

After removing duplicate search results, titles and abstracts of the identified manuscripts were screened by two authors (IMK and AMZ) to exclude studies that did not directly compare diagnostic performance of AI and HI for detection of PE on CT. We only included studies in which adjudication by separate radiologist(s) was used as the 'ground truth' to determine the performance of AI and original human interpreters. Studies that used original HI's report for assessment of AI without third party adjudication, studies that compared AI models against one another, studies comparing non-diagnostic parameters such as turn-around-times, systematic reviews, meta-analyses, and editorials were excluded at this stage. Articles regarding the use of AI for a different parameter (predictive rather than diagnostic), imaging technology (e.g. nuclear medicine scan) or pathology were also removed. Ultimately 9 studies qualified for inclusion. The screening criteria are detailed in Fig. 1.

Data extraction

Full text manuscripts of the included articles were reviewed to extract data. The number of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) classifications by AI and HI in each study were extracted. If these data were not reported, they were computed using the reported sensitivity, specificity, NPV, PPV, and accuracy. The extracted data are depicted in Table 1 [6–17]. Only the first phase of two studies [14, 15] directly compared AI vs. HI; only the data from the first phase was extracted from these manuscripts.

Statistical analysis

Müller-Peltzer K. et al. was excluded due to its false positive rate (FPR) for AI interpreters being considerably higher than that of any other study—78.4% compared to the next highest FPR of 4.2%—raising concerns that it may represent an outlier. Data from 15,963 CT scans was used for the final analysis.

Pooled estimates

The pooled estimates for sensitivity and specificity were calculated using a bivariate random-effects model implemented via the *reitsma* function from the R package

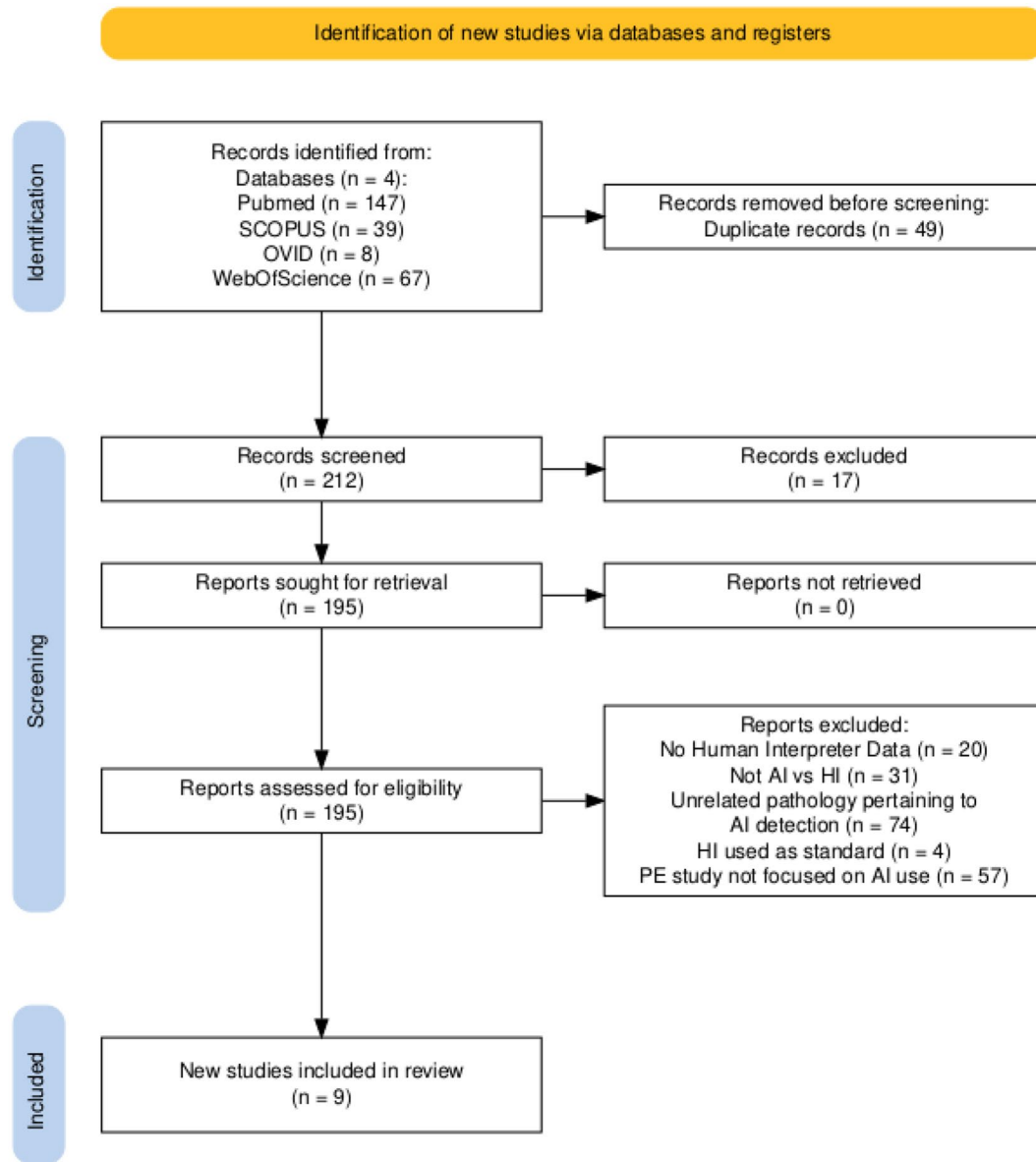


Fig. 1 PRISMA flowchart depicting study search and selection. Flowchart depicting data acquisition including the search, screening, and exclusion based on study defined criteria

Table 1 Diagnostic performance data extracted from the included studies

	AI TP	AI FP	AI FN	AI TN	HI TP	HI FP	HI FN	HI TN	TOTAL
Savage C. H. et al.	10	0	10	1447	16	1	4	1446	1467
Rothenberg S.A. et al.	46	1	27	429	64	3	9	427	503
Langius-Wiffen E. et al. 2024	64	11	3	3011	42	2	25	3020	3089
Langius-Wiffen E. et al. 2023	694	2	23	2597	657	9	60	2590	3316
Zaazoue K.A. et al.	490	4	36	961	510	9	16	956	1491
Wiklund P. et al.	68	3	7	1814	16	0	59	1817	1892
Batra K. et al.	33	5	7	2958	36	1	4	2962	3003
Cheikh A.B. et al.	176	43	14	969	171	9	19	1003	1202
Müller-Peltzer K. et al.*	118	821	64	226	180	0	2	1047	1229
TOTAL	1699	890	191	14,412	1692	34	198	15,268	17,192

*Excluded from final analysis

Table 2 Pooled estimates of the sensitivity, specificity, PPV, NPV and accuracy of AI and HI

	AI	HI	Effect size	P
Sensitivity	0.997	0.998	-0.001	0.546
Specificity	0.876	0.823	0.054	0.69
PPV	0.949	0.96	-0.011	0.727
NPV	0.99	0.986	0.004	0.589
Accuracy	0.988	0.984	0.004	0.506

mada [18, 19]. The pooled estimates for PPV and NPV were derived using the `predv_r` function, also from *mada*, by sampling from the previously estimated bivariate distribution for sensitivity and specificity. These metrics were then calculated according to the general formulas

$$PPV = \frac{Sens. \times Prev.}{Sens. \times Prev. + (1 - Spec.) \times (1 - Prev.)}$$

and

$$NPV = \frac{Spec. \times (1 - Prev.)}{Spec. \times (1 - Prev.) + Prev. \times (1 - Sens.)}$$

based on the pooled prevalence across all included studies. Likewise, the pooled estimate for accuracy was calculated according to the formula $Accuracy = Sens. \times Prev. + Spec. \times (1 - Prev.)$

using the aforementioned bivariate distribution and prevalence. To determine the pooled prevalence across all studies, a generalized linear mixed model was fitted using the `metaprop` function from the *meta* package [20].

Effect size

The estimated effect size (M) for each metric (M) was calculated as the difference between the pooled estimates of that metric for AI and human interpreters: $M = MAI - MHI$. The estimated effect size (ΔM) for each metric (M) was calculated as the difference between the pooled estimates of that metric for AI and human interpreters: $\Delta M = \widehat{M}_{AI} - \widehat{M}_{HI}$. For sensitivity and specificity, the p-value for this effect size was obtained using the `reitsma` function, with the interpreter (AI vs. HI) included as a covariate. For PPV and NPV, the p-value was determined using Welch's t-test, with standard errors for the AI and HI estimates of each metric derived from `predv_r`. For accuracy, the p-value was also determined using Welch's t-test, with the standard error derived from its sampling distribution following the same procedure implemented in `predv_r` for PPV and NPV.

Results

The pooled prevalence remained largely unchanged at 6.8% ($I^2 = 99.5\%$) without the outlier study and 7.5% ($I^2 = 99.4\%$) when all studies were included. The pooled estimates indicated that artificial intelligence (AI) and human interpreters (HI) performed similarly across all evaluated metrics (Tables 2 & 3).

Table 3 Added contribution of Aug-AI (AI - HI) to the diagnostic metrics for each included study and the pooled estimate for this contribution

	Diff Sens	Diff Spec	Diff PPV	Diff NPV	Diff Accuracy
Savage C. H. et al.	-0.300	0.001	0.059	-0.004	-0.003
Rothenberg S.A. et al.	-0.247	0.005	0.023	-0.039	-0.032
Langius-Wiffen E. et al. 2024	0.328	-0.003	-0.101	0.007	0.004
Langius-Wiffen E. et al. 2023	0.052	0.003	0.011	0.014	0.013
Zaazoue K.A. et al.	-0.038	0.005	0.009	-0.020	-0.010
Wiklund P. et al.	0.693	-0.002	-0.042	0.028	0.026
Batra K. et al.	-0.075	-0.001	-0.105	-0.001	-0.002
Cheikh A.B. et al.	0.026	-0.034	-0.146	0.004	-0.024
Müller-Peltzer K. et al.	-0.341	-0.784	-0.874	-0.219	-0.718
Pooled Estimate (p-value)	-0.001 (0.546)	0.054 (0.690)	-0.011 (0.727)	0.004 (0.589)	0.004 (0.506)

Discussion

The current meta-analysis including nearly sixteen thousand CT scans did not demonstrate any statistically significant increment or decrement in sensitivity, specificity, PPV, NPV or accuracy when studies were assessed for PE by AI and HI. Except for specificity, the pooled estimates of all diagnostic metrics were very similar for AI and HI. If the clinical question was limited to presence or absence of pulmonary arterial emboli, the current meta-analysis suggests that AI could replace HI. However, CT scan performed for PE is also meant to assess for a number of additional disease entities including but not limited to pneumonia, pulmonary edema, lung masses, pleural effusions, pneumothorax, pericardial effusion, and fractures. All of these disease entities require assessment by an HI. This means that, currently, AI can only be deployed as an augmentative tool.

A tool is of value if it augments the end user's work. From this perspective, the similar diagnostic yield of AI and HI demonstrated in the current meta-analysis suggests that the overall outcomes may not change with the addition of AI. There is a major caveat to this interpretation: pooled estimates do not delineate patient-level performance or individual outcomes in clinical practice. It is conceivable that there are some pulmonary emboli missed by HI and identified by AI and vice versa. For example, in one study [11], AI identified 5 more true positive cases than HI in aggregate. However, a closer look demonstrates that AI detected 19 cases missed by HI while HI identified 14 cases missed by AI. In such a scenario, the AI and HI would complement each other and could achieve higher diagnostic performance.

Future studies comparing AI and HI could benefit from using case-by-case outcomes to compute aggregate metrics.

Beyond basic diagnostic metrics, high performing augmentative AI can add value to radiologists' workflow by serving as a second reader to improve diagnostic confidence, aiding to achieve faster turn-around times, functioning as a QI tool to catch any missed actionable findings, and triaging studies for earlier review by the radiologist. In the last use scenario of AI-led triage, it is important to consider the potential presence of other actionable findings which are not being assessed by the AI algorithm. For example: a CT scan negative for pulmonary emboli may be positive for acute aortic injury, pneumonia, fractures, mediastinal hematoma, or active contrast extravasation – all findings that require attention sooner rather than later. In the same vein, subsegmental pulmonary embolism without right heart strain may not be as time critical as some of the aforementioned findings. Using augmentative AI to change patients' priority comes with added responsibility for the end user, especially in the acute/emergency settings. Hence, it is important to study AI's effectiveness in each distinct use scenario and to have a feed-back mechanism in place to correct and calibrate augmentative AI's performance.

Limitations

First, this study includes data from different AI algorithms. This heterogeneity is inherent to meta-analyses and appropriate statistical strategies were employed to adjust for this. The studies did not provide details about the inner workings of the AI algorithms. Such 'black box' algorithms are quite commonly seen in the AI literature. The performance of an AI algorithm depends on multiple factors (code, temperature, training datasets, validation datasets, similarity between training and real-world datasets) which vary from practice to practice. As we saw with Müller-Peltzer K. et al., one algorithm performed considerably different from the rest of the studied algorithms. The variation in the training and experience of HI is also a likely source of heterogeneity among the studies. As such, the results of the current study provide a bird's eye view of the added value of AI algorithms.

Second, we only included studies where both AI and HI read the same data and the final outcomes were adjudicated by a third party. By excluding studies that used HI as the benchmark for AI, only a subset of published data on the use of AI was included.

Third, publication bias may also limit this study's accuracy. It is likely that reports on the performance of an AI algorithm are not published until a statistically significant threshold has been achieved. As a result, studies with AI performing better are more likely to be published.

Finally, individual level data was not available. It is possible that the HI correctly read one set of studies in the

dataset while the AI correctly read another set of studies in the dataset. The case-by-case match or mismatch between the TP, FP, TN, FN metrics of AI and HI would provide a better picture of the comparative performance but could not be assessed.

Implications for future research/practice in medicine

Our findings highlight the importance of establishing the appropriate use scenario before studying AI in radiology. There is a major difference in augmentative AI (which does not assume responsibility for the outcomes) and autonomous AI (which would be liable for outcomes similar to a radiologist). Future studies of AI should assess the comparative diagnostic performance of AI and HI in both scenarios, ideally on a case-by-case basis. Other potential avenues where AI adds value such as diagnostic confidence and report turn-around-times may also be studied. In any case, it is important to explicitly define and match the intended use of AI to the study outcomes.

The radiology practices can utilize such analyses to assess and recalibrate the value proposition of any AI algorithms presented to them. Such studies can inform the choices regarding the deployment of AI as a triaging tool, a diagnostic aid, or an autonomous reader. The findings of the current studies favor the use of augmentative AI to flag studies positive for pulmonary emboli especially when radiology worklists contain copious quantities of studies. Augmentative AI could enhance workflow by reducing time-to-treatment for positive cases. Similar diagnostic performance of augmentative AI also be leveraged to enhance diagnostic confidence, although there is a risk of HI anchoring their initial correct assessment on a misclassification by AI. This will also motivate AI developers to build feedback loops to maximize augmentative AI algorithm's contribution and to eventually get to a point where AI can assume responsibility and liability for its actions.

The appropriate use scenarios will also allow the medical students considering a future in radiology to realistically assess the claims of AI replacing humans and the timeline of any such change. Premature claims regarding the autonomy of AI can lead to decreased interest in radiology and cut-off the supply line for an already undersupplied workforce in radiology.

Conclusion

A meta-analysis of augmentative AI algorithms for assessment of pulmonary emboli on CT scans demonstrated no statistically significant improvement beyond that afforded

by the human interpreters reading the same studies. Such findings can be used to readjust the use scenario of AI and to recalibrate the value proposition of augmentative AI.

Data availability The data that support the findings of this study are available on request from the corresponding author.

Declarations

Competing interests Authors confirm that we do not have competing financial or non-financial interests regarding the work studied in this paper.

References

- Center for Devices and Radiological Health. Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. U.S. Food and Drug Administration, 22 September 2021. Available at: <http://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>
- Mello-Thoms C, Mello CAB (Oct. 2023) Clinical applications of artificial intelligence in radiology. *Br J Radiol* 96(1150):20221031. <https://doi.org/10.1259/bjr.20221031>
- U.S. Food and Drug Administration. De Novo Classification Request. FDA, 8 February 2024. Accessed 11 April 2025. Available at: <https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/de-novo-classification-request>
- U.S. Food and Drug Administration. Overview of Device Regulation. FDA, 8 Feb. 2024. <https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/overview-device-regulation#510k>
- American College of Radiology (2023) Responsible for AI. *Am Coll Radiol*, <https://www.acr.org/Practice-Management-Quality-Improvement/ACR-Bulletin/Articles/January-2023/Responsible-for-AI>
- European and North American Radiology Societies (2019) Ethical considerations in AI for radiology. *Radiol Radiological Soc North Am*, <https://pubs.rsna.org/radiology/doi/full/10.1148/radio.1.2019191586>
- Haddaway NR, Matthew J, Page CC, Pritchard, McGuinness LA (2022) PRISMA2020: an R package and Shiny app for producing PRISMA 2020-Compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Syst Reviews* 18:e1230. <https://doi.org/10.1002/cl2.1230>
- The United States Food and Drug Administration (2024) Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices. FDA, Accessed 11 Dec <http://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices>
- Langius-Wiffen E, de Jong PA, Hoessein FAM, Dekker L, van den Hoven AF, Nijholt IM, Boomsma MF, Veldhuis WB (June 2023) Retrospective batch analysis to evaluate the diagnostic accuracy of a clinically deployed AI algorithm for the detection of acute pulmonary embolism on CTPA. *Insights into Imaging* 14(1):102. <https://doi.org/10.1186/s13244-023-01454-1>
- Batra K, Xi Y, Al-Hreish KM, Kay FU, Browning T, Baker C, Peshock RM (Dec. 2022) Detection of incidental pulmonary embolism on conventional Contrast-Enhanced chest CT: comparison of an artificial intelligence algorithm and clinical reports. *AJR Am J Roentgenol* 219(6):895–902. <https://doi.org/10.2214/AJR.22.27895>
- Cheikh AB, Gorincour G, Nivet H, May J, Seux M, Calame P, Thomson V, Delabrousse E, Crombé A (Sept. 2022) How artificial intelligence improves radiological interpretation in suspected pulmonary embolism. *Eur Radiol* 32(9):5831–5842. <https://doi.org/10.1007/s00330-022-08645-2>
- Langius-Wiffen E, de Jong PA, Mohamed Hoessein FA, Dekker L, van den Hoven AF, Nijholt IM, Boomsma MF, Veldhuis WB (Jan. 2024) Added value of an artificial intelligence algorithm in reducing the number of missed incidental acute pulmonary embolism in routine portal venous phase chest CT. *Eur Radiol* 34(1):367–373. <https://doi.org/10.1007/s00330-023-10029-z>
- Wiklund P, Medson K, Elf J (Feb. 2023) Incidental Pulmonary Embolism in Patients with Cancer: Prevalence, Underdiagnosis and Evaluation of an AI Algorithm for Automatic Detection of Pulmonary Embolism. *Eur Radiol* 33(2):1185–1193. <https://doi.org/10.1007/s00330-022-09071-0>
- Rothenberg SA, Savage CH, Abou Elkassem A, Singh S, Abozeed M, Hamki O, Junck K et al (Oct. 2023) Prospective evaluation of AI triage of pulmonary emboli on CT pulmonary angiograms. *Radiology* 309(1):e230702. <https://doi.org/10.1148/radiol.230702>
- Savage CH, Elkassem AA, Hamki O, Sturdivant A, Benson D, Grumley S, Tzabari J et al (Sept. 2024) Prospective evaluation of artificial intelligence triage of incidental pulmonary emboli on Contrast-Enhanced CT examinations of the chest or abdomen. *AJR Am J Roentgenol* 223(3):e2431067. <https://doi.org/10.2214/AJR.24.31067>
- Zaazoue KA, McCann MR, Ahmed K, Ahmed IO, Cortopassi YM, Erben, Brian P, Little JT, Stowell et al (June. 2023) Evaluating the Performance of a Commercially Available Artificial Intelligence Algorithm for Automated Detection of Pulmonary Embolism on Contrast-Enhanced Computed Tomography and Computed Tomography Pulmonary Angiography in Patients with Coronavirus Disease 2019. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 7(3):143–152. <https://doi.org/10.1016/j.mayocpiqo.2023.03.001>
- Müller-Peltzer K, Kretzschmar L, de Figueiredo GNegrão, Crispin A, Stahl R, Bamberg F, Christoph G (Dec. 2021) Trumm. Present Limitations of Artificial Intelligence in the Emergency Setting - Performance Study of a Commercial, Computer-Aided Detection Algorithm for Pulmonary Embolism. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgebenden Verfahren* 193(12):1436–1444. <https://doi.org/10.1055/a-1515-2923>
- Reitsma JB, Adrian S, Glas, Anne WS, Rutjes, Rutger JPM, Scholten PM, Bossuyt, Zwinderman AH (Oct. 2005) Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 58(10):982–990. <https://doi.org/10.1016/j.jclinepi.2005.02.022>
- CRAN: Package Mada. R-Project.org (2022) Accessed 11 Dec. 2024 <http://cran.r-project.org/package=mada>
- Doebler P (2022) Meta-Analysis of Diagnostic Accuracy (Version 0.5.11) [Software]. <http://cran.r-project.org/web/packages/mada/mada.pdf>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.