The next generation GBCA from Guerbet is here
Explore new possibilities
© Guerbet 2024 GUOB220151-A
Guerbet

AJNR

AI Efficacy as a Function of Trainee Interpreter Proficiency: Lessons from a Randomized Controlled Trial

David A. Fussell, Cynthia C. Tang, Jake Sternhagen, Varun V. Marrey, Kelsey M. Roman, Jeremy Johnson, Michael J. Head, Hayden R. Troutt, Charles H. Li, Peter D. Chang, John Joseph and Daniel S. Chow

This information is current as of July 16, 2024.

AJNR Am J Neuroradiol published online 21 June 2024
http://www.ajnr.org/content/early/2024/06/21/ajnr.A8387

Published June 21, 2024 as 10.3174/ajnr.A8387
This preprint represents the accepted version of the article and also includes the supplemental material; it differs from the printed version of the article.

ORIGINAL RESEARCH

AI Efficacy as a Function of Trainee Interpreter Proficiency: Lessons from a Randomized Controlled Trial

David A. Fussell, Cynthia C. Tang, Jake Sternhagen, Varun V. Marrey, Kelsey M. Roman, Jeremy Johnson, Michael J. Head, Hayden R. Troutt, Charles H. Li, Peter D. Chang, John Joseph, Daniel S. Chow

ABSTRACT

BACKGROUND AND PURPOSE: Recently, AI tools have been deployed with increasing speed in educational and clinical settings. However, the use of AI by trainees across different levels of experience has not been well studied. This study investigates the impact of AI assistance on diagnostic accuracy for intracranial hemorrhage (ICH) and large vessel occlusion (LVO) by medical students (MS) and resident trainees (RT).

MATERIALS AND METHODS: This prospective study was conducted between March 2023 and October 2023. MS and RT were asked to identify ICH and LVO in 100 non-contrast head CTs and 100 head CTAs, respectively. One group received diagnostic aid simulating AI for ICH only (n = 26), the other for LVO only (n = 28). Primary outcomes included accuracy, sensitivity, and specificity for ICH / LVO detection without and with aid. Study interpretation time was a secondary outcome. Individual responses were pooled and analyzed with chi-square; differences in continuous variables were assessed with ANOVA.

RESULTS: 48 participants completed the study, generating 10,779 ICH or LVO interpretations. With diagnostic aid, MS accuracy improved 11.0 points ($P < .001$) and RT accuracy showed no significant change. ICH interpretation time increased with diagnostic aid for both groups ($P < .001$) while LVO interpretation time decreased for MS ($P < .001$). Despite worse performance in detection of the smallest vs. the largest hemorrhages at baseline, MS were not more likely to accept a true positive AI result for these more difficult tasks. Both groups were considerably less accurate when disagreeing with the AI or when supplied with an incorrect AI result.

CONCLUSIONS: This study demonstrated greater improvement in diagnostic accuracy with AI for MS compared to RT. However, MS were less likely than RT to overrule incorrect AI interpretations and were less accurate, even with diagnostic aid, than the AI was by itself.

ABBREVIATIONS: ICH = intracranial hemorrhage; LVO = large vessel occlusion; MS = medical students; RT = resident trainees.

Received month day, year; accepted after revision month day, year.
From the Department of Radiological Sciences, University of California, Irvine, Irvine, CA, USA (D.A.F., C.C.T., J.S., V.V.M., J.J., H.R.T., C.H.L., P.D.C., D.S.C); School of Medicine, University of California, Irvine, Irvine, CA, USA (M.J.H.); Paul Merage School of Business, University of California, Irvine, Irvine, CA, USA (J.J.).

Author Daniel S. Chow is the recipient of grants from the Norris Foundation and consulting fees from Canon Medical and has patents planned, issued, or pending with University of California, Irvine.

Please address correspondence to David Fussell, MD, Department of Radiological Sciences, University of California Irvine Medical Center, 101 The City Drive South, Orange, CA 92868, fusselld@hs.uci.edu.

SUMMARY SECTION

PREVIOUS LITERATURE: Prior work suggests that physicians with less experience in radiology benefit the most from AI-assistance, while a recent large-scale study found that experience-based factors do not reliably predict the impact of AI assistance. The factors influencing use and trust across different levels of interpreter expertise remain poorly understood.

KEY FINDINGS: Diagnostic aid simulating AI demonstrated improvement in ICH and LVO detection for medical students, but not for resident trainees. Furthermore, MS were less likely than RT to overrule incorrect aid interpretations and were less accurate than the simulated AI alone.

KNOWLEDGE ADVANCEMENT: AI may provide a greater benefit for non-experts; however, a threshold level of experience may be necessary for the safe and effective use of deep learning tools.

INTRODUCTION

Over the last several decades, the volume of medical imaging has dramatically increased within the United States healthcare system.[1,2] Drivers of high volume include increasing population size and age, growing emphasis on cross-sectional studies, and a lack of widespread adoption of evidence-based guidelines for imaging utilization.[3] Although imaging is intended to improve medical decision making, increased imaging volume demands increased throughput from radiologists, which increases the risk of diagnostic error; this may have devastating consequences for patient care.[4,5] Moreover, medical error is expensive, accounting for an estimated $17 billion to $29 billion in annual excess spending in the United States.[6]

Copyright 2024 by American Society of Neuroradiology.
Copyright 2024 by American Society of Neuroradiology.

More recently, there has been an exponential increase in the number of available artificial intelligence (AI) products, which represent one solution for managing high study volumes. Several studies have demonstrated that these tools enhance physician performance and may prevent burnout by reducing reading time and improving diagnostic accuracy.[7–11] AI is increasingly used to support clinical decision making and to triage acute findings. In a recent randomized clinical trial of 443 participants across four comprehensive stroke centers, Gutierrez et al. showed significantly reduced time to endovascular thrombectomy for patients with large vessel occlusion (LVO) using an LVO detection AI algorithm that automatically alerts clinicians and radiologists.[12] Although machine learning has demonstrated impressive performance in detecting specific imaging abnormalities, current technology is limited to simple tasks, lacks clinical decision-making capabilities, and continues to require physician oversight.[13]

The increasing prevalence of AI in radiology raises questions about its role in medical education and resident training. As many as 40% of imaging studies from teaching institutions are cosigned by radiology trainees.[14,15] Although several studies have reported improved trainee performance with deep learning tools, the factors influencing use and trust across different levels of interpreter expertise remain poorly understood.[16]

The purpose of this randomized, controlled trial is to investigate how having an AI result available at the time of interpretation influences accuracy and interpretation time across different levels of medical training and task complexity. We hypothesize that such diagnostic aid will increase accuracy and decrease interpretation time for all trainees, but that the effect will be greater for less experienced readers. Similarly, we expect the benefit to be greater for tasks of greater complexity. The study will also investigate whether level of training influences how trainees deal with incorrect diagnostic aid. This article follows the CONSORT reporting guidelines.

MATERIALS AND METHODS

Study Design

This prospective study was conducted at the University of California, Irvine and approved by our institutional review board. After providing written informed consent, medical students (MS) and resident trainees (RT) were randomized to one of two groups: (A) ICH detection without diagnostic aid and LVO detection with diagnostic aid or (B) ICH detection with diagnostic aid and LVO detection without diagnostic aid. The primary interpretation target of LVO detection was identification of occlusions in the M1 segment of the middle cerebral artery (MCA). Randomization and intervention assignment were performed following a 1:1 allocation ratio. To limit the potential for study participants to assess the fixed accuracy of the provided diagnostic aid, positive and negative cases were presented in a random sequence and false positive / false negative diagnostic aid responses were randomly distributed. All medical students attended a 60-minute lecture on to the fundamentals of recognizing ICH and LVO in CT scans through neuroanatomy and case examples.

Participants

The MS group consisted of first- and second-year medical students from the University of California. RT consisted of U.C. Irvine radiology residents in their third to fifth post-graduate year. Recruitment occurred between January 2023 and October 2023. Participants who did not complete both assigned tasks were excluded. Participants did not know the accuracy of the AI beforehand.

Viewer

Participants were tasked with completing two reading sessions: (1) 100 non-contrast head CTs and (2) 100 CT angiographies of the head. Both sets were balanced (50:50) between normal and abnormal (presence / absence of ICH or LVO). Diagnostic aid was shown to participants as a binary yes / no for the presence of ICH or LVO. Tasks were completed on participants' devices using an established, research-grade viewing platform offering standard functionality such as zoom and adjustable window / level. Responses were collected in a separate browser window. To ensure a robust set of false positive / false negative aid responses, diagnostic aid was calibrated to have both sensitivity and specificity of 80%.

Dataset

The dataset used for this study included 200 total de-identified CT scans: 50 CT angiographies with LVO, 50 non-contrast head CTs with ICH, and 100 CT angiographies and non-contrast head CT scans with no pathology. The same scans were used for sessions in which participants had or did not have access to diagnostic aid. 200 patients were included in the dataset.

Ground Truth Definition

Ground truth was established by an experienced neuroradiologist (D.C., 12 years of experience).

Outcome Measures

The primary outcome measures included reader accuracy, sensitivity, and specificity without or with diagnostic aid. These were determined according to whether the participant's answer ("yes" or "no" to the presence of ICH or LVO) agreed with the ground truth. The secondary outcome measure was interpretation time, which was calculated automatically for each case using Qualtrics survey software.

Subgroup Evaluation

Primary outcomes were evaluated in several subgroups: within tasks (ICH and LVO); according to whether the user agreed or disagreed with the diagnostic aid interpretation; and according to whether the supplied aid interpretation was accurate. After being segmented with

a previously validated algorithm, positive ICH cases were split into quintiles according to hemorrhage size and primary outcome measures were assessed within these quintiles.

Statistical Analysis

Based on an anticipated diagnostic accuracy of 70% for MS and 75% for RT, a desired power of 80%, and expected higher enrollment rate for MS participants, we estimated that a total of at least 4,000 responses, or 20 participants, would be required in each arm. Statistical analyses were performed using software (Python version 3.10, the Python Software Foundation; Pandas version 2.1.0, The Pandas Development Team). De-identified user and response information was stored as raw data within Microsoft Excel. Answers with response times greater than 4 standard deviations above the mean were discarded. Mean accuracy, sensitivity, and specificity were computed for each participant and used as data points. Mean accuracy, sensitivity, and specificity were computed for each group and compared using a t-test. A t-test was also used to compare median response times. ANOVA was used for comparison of accuracy across different hemorrhage sizes in the ICH task.

RESULTS

Participants

A total of 93 participants expressed interest in the study and began the consent process. 68 participants provided written informed consent and were enrolled. Ultimately, 48 participants completed the study (Figure 1). This group included 37 MS and 11 RT. The final MS group included first and second year medical students recruited from University of California Irvine and Riverside medical schools. The RT group included 11 UCI radiology residents. Given that sample size, the minimum difference in accuracy for RT without and with diagnostic aid that could have been detected at 80% power was 8.5 percent.

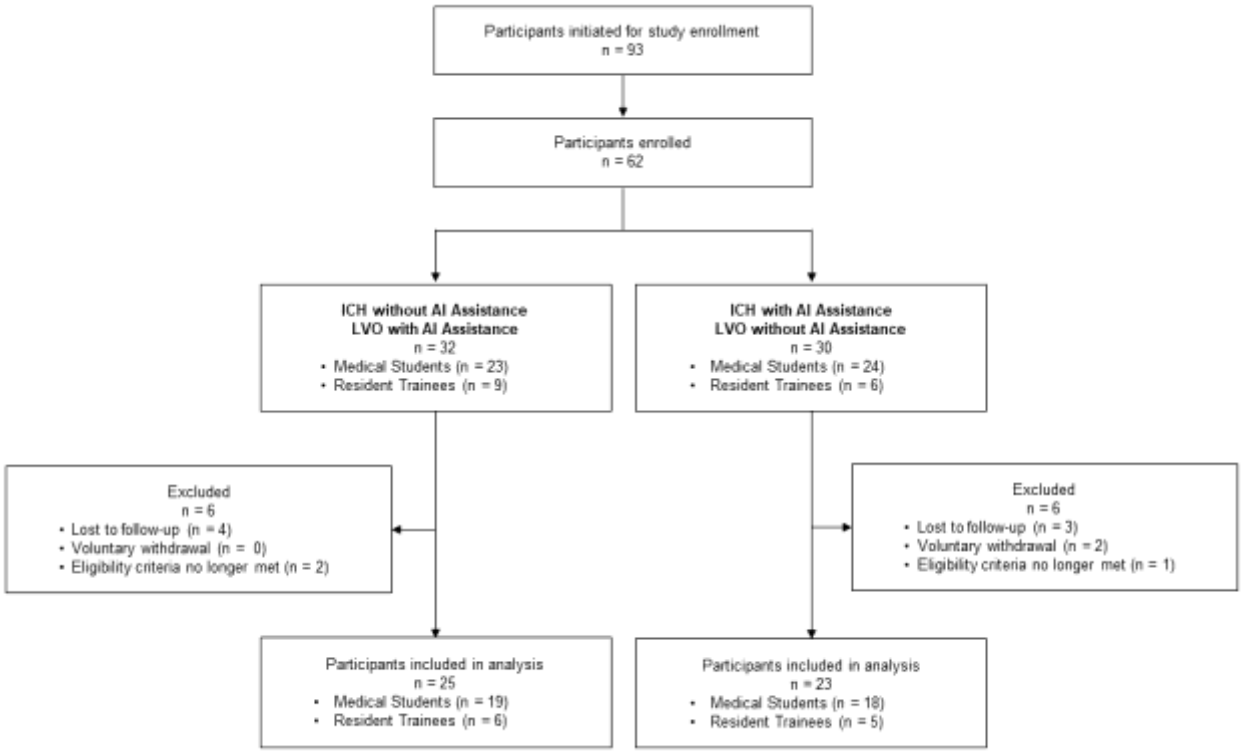

FIG 1. Participant enrollment flowchart.

Primary Analysis

With diagnostic aid, MS accuracy improved 11.0 points (62.6% to 73.6%, P < .001, Figure 2), while RT accuracy showed no significant change. MS sensitivity improved from 48.0% to 68.6% with aid (P < .001, Table 1), while specificity was not significantly different. For RT, sensitivity improved from 74.0% to 86.0% (P = .025). Specificity was not significantly different.

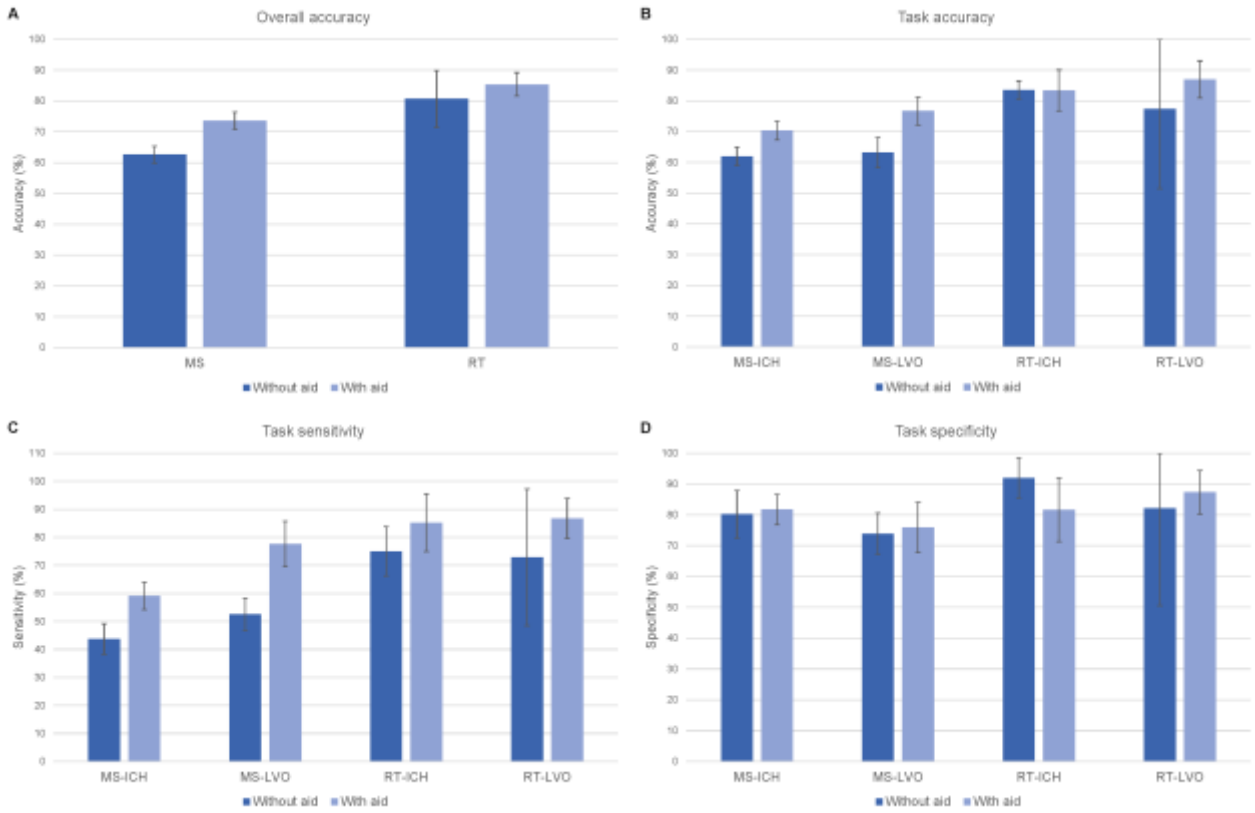

FIG 2. (A) Overall accuracy for MS and RT without and with diagnostic aid. (B – D) Accuracy, sensitivity, and specificity changes within each task without and with diagnostic aid. MS: medical students; RT: resident trainees; ICH: intracranial hemorrhage; LVO: large vessel occlusion.

Task Analysis

Next, we assessed differences in benefit of diagnostic aid across different tasks. For the ICH task, MS accuracy improved from 62.0% to 70.4% (P < .001, Table 1) with aid. On the LVO task, MS accuracy improved from 63.2% to 76.7% (P < .001).

Table 1: Accuracy, sensitivity, and specificity overall and for each individual task, without and with diagnostic aid.

	Metric	Without aid (mean [95% CI])	With aid	P-value
MS Overall	Accuracy	62.6 [59.9 - 65.3]	73.6 [70.8 - 76.5]	< .001
	Sensitivity	48.0 [43.9 - 52.0]	68.6 [63.0 - 74.2]	< .001
	Specificity	77.2 [72.2 - 82.2]	78.7 [74.0 - 83.3]	0.65
RT Overall	Accuracy	80.7 [71.5 - 90.0]	85.4 [81.6 - 89.1]	0.32
	Sensitivity	74.0 [64.7 - 83.3]	86.0 [80.1 - 91.9]	0.025
	Specificity	87.5 [75.8 - 99.2]	84.7 [79.5 - 89.9]	0.63
MS-ICH	Accuracy	62.0 [59.0 - 64.9]	70.4 [67.3 - 73.5]	< .001
	Sensitivity	43.7 [38.2 - 49.2]	59.1 [50.9 - 67.3]	0.002
	Specificity	80.2 [72.5 - 87.9]	81.7 [76.8 - 86.5]	0.74
MS-LVO	Accuracy	63.2 [58.3 - 68.1]	76.7 [72.2 - 81.2]	< .001
	Sensitivity	52.5 [46.7 - 58.3]	77.6 [72.1 - 83.0]	< .001
	Specificity	73.9 [67.2 - 80.7]	75.9 [67.8 - 84.0]	0.70
RT-ICH	Accuracy	83.5 [80.6 - 86.4]	83.4 [76.5 - 90.3]	0.97
	Sensitivity	75.0 [66.2 - 83.8]	85.2 [74.4 - 96.0]	0.079
	Specificity	92.0 [85.5 - 98.5]	81.6 [71.3 - 91.9]	0.041
RT-LVO	Accuracy	77.4 [51.5 - 103.4]	87.0 [81.0 - 92.9]	0.31
	Sensitivity	72.8 [48.4 - 97.2]	86.7 [76.5 - 96.8]	0.16
	Specificity	82.1 [50.5 - 113.7]	87.3 [80.2 - 94.3]	0.64

Note: MS: medical students; RT: resident trainees; ICH: intracranial hemorrhage; LVO: large vessel occlusion.

For RT performing the ICH task, accuracy and sensitivity were not significantly changed with diagnostic aid. Specificity actually decreased, 92.0 to 81.6% (P = .041, Table 1). In the LVO task, RT accuracy, sensitivity, and specificity were not significantly changed.

Within the ICH task, we hypothesized that diagnostic aid would be more helpful in the detection of smaller hemorrhages. To assess this, we segmented positive ICH cases and split hemorrhages into quintiles according to size. For MS, mean accuracies without aid were significantly different across hemorrhage sizes (ranging from 21.1% for the smallest hemorrhages to 75.8% for the largest hemorrhages, ANOVA P < .001; Figure 3). For all but the largest hemorrhages, accuracy improvement with aid was statistically significant (P < .05).

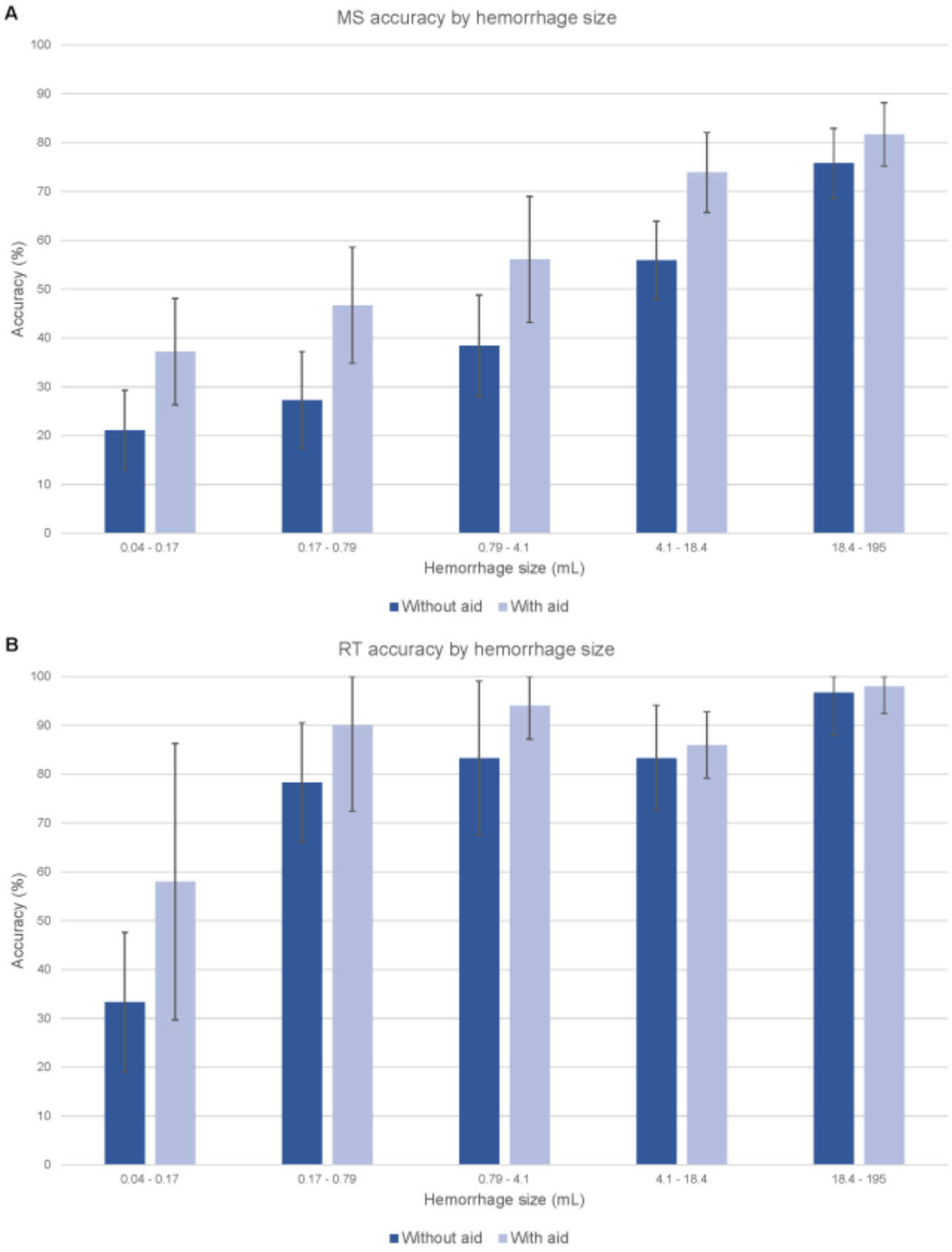

FIG 3. Overall accuracy for (A) MS and (B) RT without and with diagnostic aid across different volumes of intracranial hemorrhage. MS: medical students; RT: resident trainees; ICH: intracranial hemorrhage; LVO: large vessel occlusion.

For RT, mean accuracies without aid ranged from 33.3% for very small hemorrhages to 96.7% for the largest hemorrhages (ANOVA $P < .001$, Figure 3). The accuracy benefit conferred by diagnostic aid was not statistically significant within any quintile and did not vary significantly across hemorrhage sizes.

Table 2: Median interpretation time by AI response type.

Level	Task	AI TP	AI TN	AI FP	AI FN	P-value
MS	ICH	23.5	23.2	29.3	21.2	0.04
MS	LVO	24.4	21.2	22.9	24.4	.007
RT	ICH	21.1	35.1	40.4	25.7	< .001
RT	LVO	31.8	33.4	43.2	33.3	0.69

(P < .001). For RT performing the ICH task, accuracy dropped from 87.5% to 67.0% with incorrect aid response (P < .001). On the LVO task, accuracy dropped from 91.7% to 68.2% (P < .001).

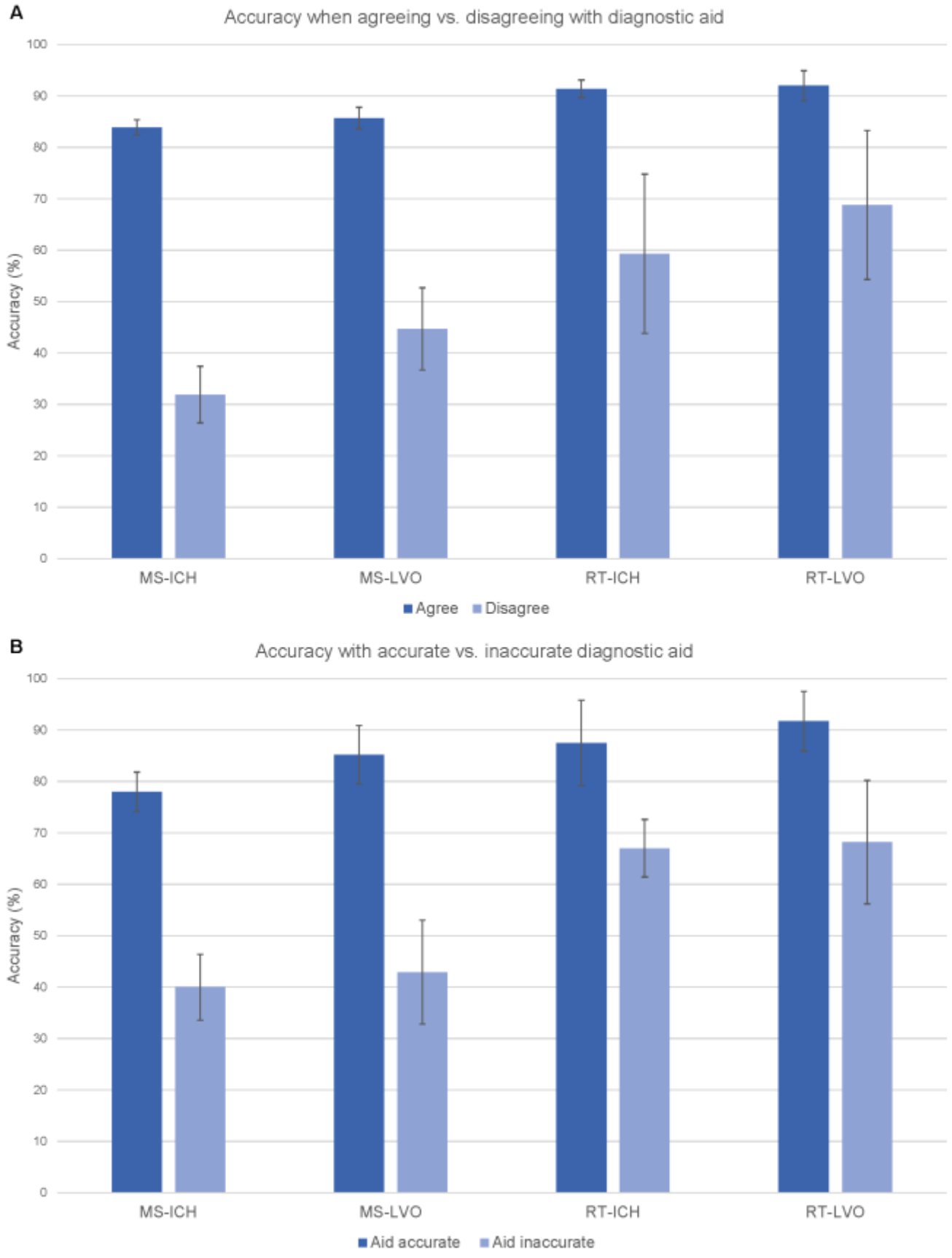

FIG 5. (A) Accuracy for each task when agreeing vs. disagreeing with the diagnostic aid. (B) Accuracy for each task when the diagnostic aid interpretation was accurate vs. inaccurate. All differences were statistically significant (P < .05). MS, medical student; RT, resident trainee; ICH, intracranial hemorrhage; LVO, large vessel occlusion.

DISCUSSION

Prior work suggests that physicians with less experience in radiology benefit the most from AI-assistance, while a recent large-scale study found that experience-based factors do not reliably predict the impact of AI assistance. [17,18] Our study, using diagnostic aid to simulate AI, demonstrates significant increase in accuracy with diagnostic aid for MS but no significant increase for RT. When ICH evaluations were stratified by hemorrhage size, both MS and RT were less accurate at baseline in detecting the smallest vs. the largest hemorrhages (MS, 21.1% vs. 75.8%; RT, 33.3% vs 96.7%, both $P < .001$). However, the benefit conferred by diagnostic aid did not vary significantly across hemorrhage sizes. Diagnostic aid had no statistically significant effects on interpretation time. Both groups were significantly less accurate when disagreeing with the aid interpretation and when supplied with an incorrect aid response. MS, but not RT, were less accurate, even with diagnostic aid, than the simulated AI was by itself.

For both MS and RT, essentially all the benefit of diagnostic aid came from increased sensitivity. This is concordant with prior studies that have demonstrated a greater improvement in sensitivity with AI assistance among radiologists.[19–21] It may be more difficult for AI assistance to increase specificity, as this would require users to abandon an initial positive read in favor of the true negative AI response, which by nature could not be supported by a discrete finding in the scan. On the other hand, a user considering a true positive AI result might, on second look, identify the finding that triggered the AI result and more readily change their initial response.

We expected that baseline performance would be lower for more complex tasks and that diagnostic aid would offer greater benefit in these situations. Given that the difficulty of ICH detection depends on the size of the ICH, we split positive cases into quintiles by hemorrhage volume. As expected, both MS and RT demonstrated worse baseline performance in detecting smaller hemorrhages. However, the benefit of diagnostic aid for MS was similar across the smallest 4 quintiles of hemorrhage. This means that for the smallest hemorrhages, MS were more likely to disregard a true positive aid response. This may have been due to anchoring bias, where a participant remains fixed to an initial diagnostic interpretation despite being provided with new data suggesting an alternative; this has been shown to be a significant bias in radiology.[22,23] One strategy for overcoming this bias might be to use diagnostic aids that explicitly identify the suspected abnormality. Currently, AI triage tools are prohibited by FDA regulations from annotating diagnostic images in any way, but annotation may be an important consideration in the implementation of future AI systems.

We expected that interpretation time would decrease across all tasks when a diagnostic aid was available (as seen in a recent prospective study[24]), but the actual effects were mixed and not statistically significant. For the ICH task, read times were greatest when the diagnostic aid response was a false positive; this may reflect time spent searching the entire brain volume for a finding that might have triggered the response. This would be less of an issue on the LVO task, which focused on a small anatomic area around the proximal MCAs. Again, it is likely that an AI system highlighting a suspected abnormality, in addition to providing a categorical result, would show a more robust decrease in read times across tasks.

A recent study by Yu et al.[18] demonstrated that, contrary to what one might expect, less experienced board-certified radiologists did not benefit more from AI assistance than more experienced radiologists, and that overall AI benefit was small. Gaube et al.[25] demonstrated a significant improvement in accuracy for non-expert physicians (internal or emergency medicine) but not for radiologists with AI assistance. Our study demonstrated a significant benefit from diagnostic aid for MS, but not RT. Although the study designs and specific diagnostic tasks investigated are different, these results suggest that, as the experience of the user increases, the relationship between AI assistance and accuracy becomes more complicated. However, in our study, despite clearly benefitting from diagnostic aid, MS were still less accurate, even with aid, than the simulated AI was by itself. This may be a manifestation of the Dunning-Krueger effect, a cognitive bias where subjects overestimate their ability to perform a task, despite having limited task-specific expertise.[26]

Our results have several implications for the clinical implementation of AI, particularly in an educational setting. Although AI assistance appears to be of greater benefit to trainees, given that MS with diagnostic aid were less accurate than the simulated AI itself, there may be a minimum threshold of competency required to use radiology AI tools safely and effectively. Our results further demonstrated that the primary benefit of diagnostic aid to MS and RT was to increase sensitivity, without decreasing specificity. If trainees are more likely to be influenced by a true positive AI response than by a false negative one, future AI algorithms might be most beneficial if calibrated to have high sensitivity, even at the expense of decreased specificity. This would also accord with the perspective that, for example when interpreting ER studies overnight, it is more costly for trainees to miss a real positive finding than to imagine one that isn't actually there.

Our study had limitations. Different groups completed each task without or with diagnostic aid, and metrics to establish baseline proficiency were not available, so that individual user competence might have affected differences in accuracy. Our RT group was also relatively small, limiting the resolution of the study to detect differences in accuracy without and with diagnostic aid. The simulated diagnostic aid did not provide visual depictions of the suspected abnormalities, though we note that current FDA rules prohibit triage applications from marking up diagnostic images in any way. Additionally, the accuracy, specificity, and sensitivity of the simulated AI were fixed at 80%, which is low compared to currently available tools; however, our results may provide a baseline against which future studies can assess the impact of a more accurate AI. Finally, the study was not conducted during routine clinical practice using a standard PACS, which could affect the generalizability of the results. Future work is needed to study the integration of AI assistance into clinical workflow and to assess the effects of different baseline AI accuracies.

CONCLUSIONS

This study demonstrated improvement in ICH and LVO detection with simulated AI for MS, but not for RT, suggesting that AI may provide a greater benefit for non-experts. However, MS were less likely than RT to overrule incorrect aid interpretations, and in fact were less accurate than the simulated AI alone, suggesting that a threshold level of experience may be necessary for the safe and effective use of deep learning tools. To aid in optimal deployment of AI in the educational setting, future work should include additional participants from other institutions at different levels of experience as well as investigating different methods of reporting AI results.

ACKNOWLEDGMENTS

This study is funded by the Kenneth T. & Eileen L. Norris Foundation. Statistical analyses were performed by author DF.

REFERENCES

1. Chartrand G, Cheng PM, Vorontsov E, et al. Deep Learning: A Primer for Radiologists. *RadioGraphics* 2017;37:2113–31.

2. Heit JJ, Iv M, Wintermark M. Imaging of Intracranial Hemorrhage. *J Stroke* 2017;19:11–27.

3. Fasen B a. CM, Heijboer RJJ, Hulsmans F-JH, et al. CT Angiography in Evaluating Large-Vessel Occlusion in Acute Anterior Circulation Ischemic Stroke: Factors Associated with Diagnostic Error in Clinical Practice. *American Journal of Neuroradiology* 2020;41:607–11.

4. Matsoukas S, Scaggiante J, Schuldt BR, et al. Accuracy of artificial intelligence for the detection of intracranial hemorrhage and chronic cerebral microbleeds: a systematic review and pooled analysis. *Radiol Med* 2022;127:1106–23.

5. Rava RA, Seymour SE, LaQue ME, et al. Assessment of an Artificial Intelligence Algorithm for Detection of Intracranial Hemorrhage. *World Neurosurg* 2021;150:e209–17.

6. Petry M, Lansky C, Chodakiewitz Y, et al. Decreased Hospital Length of Stay for ICH and PE after Adoption of an Artificial Intelligence-Augmented Radiological Worklist Triage System. *Radiol Res Pract* 2022;2022:2141839.

7. Pinto Dos Santos D, Giese D, Brodehl S, et al. Medical students' attitude towards artificial intelligence: a multicentre survey. *Eur Radiol* 2019;29:1640–6.

8. Yang L, Ene IC, Arabi Belaghi R, et al. Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. *Eur Radiol* 2022;32:1477–95.

9. Juravle G, Boudouraki A, Terziyska M, et al. Trust in artificial intelligence for medical diagnoses. *Prog Brain Res* 2020;253:263–82.

10. Wagner AR, Borenstein J, Howard A. Overtrust in the robotic age. *Commun ACM* 2018;61:22–4.

11. Borracci RA, Arribalzaga EB. The Incidence of Overconfidence and Underconfidence Effects in Medical Student Examinations. *J Surg Educ* 2018;75:1223–9.

12. Martinez-Gutierrez JC, Kim Y, Salazar-Marioni S, et al. Automated Large Vessel Occlusion Detection Software and Thrombectomy Treatment Times: A Cluster Randomized Clinical Trial. *JAMA Neurology* 2023;80:1182–90.

13. Skitka LJ, Mosier KL, Burdick M, et al. Automation bias and errors: are crews better than individuals? *Int J Aviat Psychol* 2000;10:85–97.

14. Itoh M. Toward overtrust-free advanced driver assistance systems. *Cogn Tech Work* 2012;14:51–60.

15. Kapoor N, Gaviola G, Wang A, et al. Quantifying and Characterizing Trainee Participation in a Major Academic Radiology Department. *Curr Probl Diagn Radiol* 2019;48:436–40.

16. Arthur Jr. W, Bennett Jr. W, Stanush PL, et al. Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis. *Human Performance* 1998;11:57–101.

17. Li D, Pehrson LM, Lauridsen CA, et al. The Added Effect of Artificial Intelligence on Physicians' Performance in Detecting Thoracic Pathologies on CT and Chest X-ray: A Systematic Review. *Diagnostics (Basel)* 2021;11:2206.

18. Yu F, Moehring A, Banerjee O, et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat Med* 2024;30:837–49.

19. Jacques T, Cardot N, Ventre J, et al. Commercially-available AI algorithm improves radiologists' sensitivity for wrist and hand fracture detection on X-ray, compared to a CT-based ground truth. *Eur Radiol* https://doi.org/10.1007/s00330-023-10380-1.

20. Watanabe Y, Tanaka T, Nishida A, et al. Improvement of the diagnostic accuracy for intracranial haemorrhage using deep learning-based computer-assisted detection. *Neuroradiology* 2021;63:713–20.

21. Ewals LJS, van der Wulp K, van den Borne BEEM, et al. The Effects of Artificial Intelligence Assistance on the Radiologists' Assessment of Lung Nodules on CT Scans: A Systematic Review. *Journal of Clinical Medicine* 2023;12:3536.

22. Busby LP, Courtier JL, Glastonbury CM. Bias in Radiology: The How and Why of Misses and Misinterpretations. *Radiographics* 2018;38:236–47.

23. Lee CS, Nagy PG, Weaver SJ, et al. Cognitive and system factors contributing to diagnostic errors in radiology. *AJR Am J Roentgenol* 2013;201:611–7.

24. Yacoub B, Varga-Szemes A, Schoepf UJ, et al. Impact of Artificial Intelligence Assistance on Chest CT Interpretation Times: A Prospective Randomized Study. *American Journal of Roentgenology* 2022;219:743–51.

25. Gaube, S., Suresh, H., Raue, M. et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci Rep* 13, 1383 (2023).

26. Kruger J, Dunning D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 1999;77:1121–34.