## RSNA Press Release

# Incorrect AI Advice Influences Diagnostic Decisions

Released: November 19, 2024

OAK BROOK, Ill. — When making diagnostic decisions, radiologists and other physicians may rely too much on artificial intelligence (AI) when it points out a specific area of interest in an X-ray, according to a study published today in *Radiology*, a journal of the Radiological Society of North America (RSNA).

"As of 2022, 190 radiology AI software programs were approved by the U.S. Food and Drug Administration," said one of the study's senior authors, Paul H. Yi, M.D., director of intelligent imaging informatics and associate member in the Department of Radiology at St. Jude Children's Research Hospital in Memphis, Tennessee. "However, a gap between AI proof-of-concept and its real-world clinical use has emerged. To bridge this gap, fostering appropriate trust in AI advice is paramount."

In the multi-site, prospective study, 220 radiologists and internal medicine/emergency medicine physicians (132 radiologists) read chest X-rays alongside AI advice. Each physician was tasked with evaluating eight chest X-ray cases alongside suggestions from a simulated AI assistant with diagnostic performance comparable to that of experts in the field. The clinical vignettes offered frontal and, if available, corresponding lateral chest X-ray images obtained from Beth Israel Deaconess Hospital in Boston via the open-source MIMI Chest X-Ray Database. A panel of radiologists selected the set of cases that simulated real-world clinical practice.

download photo

Paul H. Yi, M.D.

For each case, participants were presented with the patient's clinical history, the AI advice and X-ray images. AI provided either a correct or incorrect diagnosis with local or global explanations. In a local explanation, AI highlights parts of the image deemed most important. For global explanations, AI provides similar images from previous cases to show how it arrived at its diagnosis.

"These local explanations directly guide the physician to the area of concern in real-time," Dr. Yi said. "In our study, the AI literally put a box around areas of pneumonia or other abnormalities."

The reviewers could accept, modify or reject the AI suggestions. They were also asked to report their confidence level in the findings and impressions and to rank the usefulness of the AI advice.

Using mixed-effects models, study co-first authors Drew Prinster, M.S., and Amama Mahmood, M.S., computer science Ph.D. students at Johns Hopkins University in Baltimore, led the researchers in analyzing the effects of the experimental variables on diagnostic accuracy, efficiency, physician perception of AI usefulness, and "simple trust" (how quickly a user agreed or disagreed with AI advice). The researchers controlled for factors like user demographics and professional experience.

The results showed that reviewers were more likely to align their diagnostic decision with AI advice and underwent a shorter period of consideration when AI provided local explanations.

"Compared with global AI explanations, local explanations yielded better physician diagnostic accuracy when the AI advice was correct," Dr. Yi said. "They also increased diagnostic efficiency overall by reducing the time spent considering AI advice."

When the AI advice was correct, the average diagnostic accuracy among reviewers was 92.8% with local explanations and 85.3% with global explanations. When AI advice was incorrect, physician accuracy was 23.6% with local and 26.1% with global explanations.

"When provided local explanations, both radiologists and non-radiologists in the study tended to trust the AI diagnosis more quickly, regardless of the accuracy of AI advice," Dr. Yi said.

Study co-senior author, Chien-Ming Huang, Ph.D., John C. Malone Assistant Professor in the Department of Computer Science at Johns Hopkins University, pointed out that this trust in AI could be a double-edged sword because it risks over-reliance or automation bias.

"When we rely too much on whatever the computer tells us, that's a problem, because AI is not always right," Dr. Yi said. "I think as radiologists using AI, we need to be aware of these pitfalls and stay mindful of our diagnostic patterns and training."

Based on the study, Dr. Yi said AI system developers should carefully consider how different forms of AI explanation might impact reliance on AI advice.

"I really think collaboration between industry and health care researchers is key," he said. "I hope this paper starts a dialog and fruitful future research collaborations."

"Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI." Collaborating with Dr. Yi, Dr. Huang, Prinster and Mahmood were Suchi Saria, Ph.D., Jean Jeudy, M.D., and Cheng Ting Lin, M.D.

*Radiology* is edited by Linda Moy, M.D., New York University, New York, N.Y., and owned and published by the Radiological Society of North America, Inc. (https://pubs.rsna.org/journal/radiology)

RSNA is an association of radiologists, radiation oncologists, medical physicists and related scientists promoting excellence in patient care and health care

delivery through education, research and technologic innovation. The Society is based in Oak Brook, Illinois. (RSNA.org)

For patient-friendly information on chest X-rays, visit *RadiologyInfo.org*.

Images (JPG, TIF):

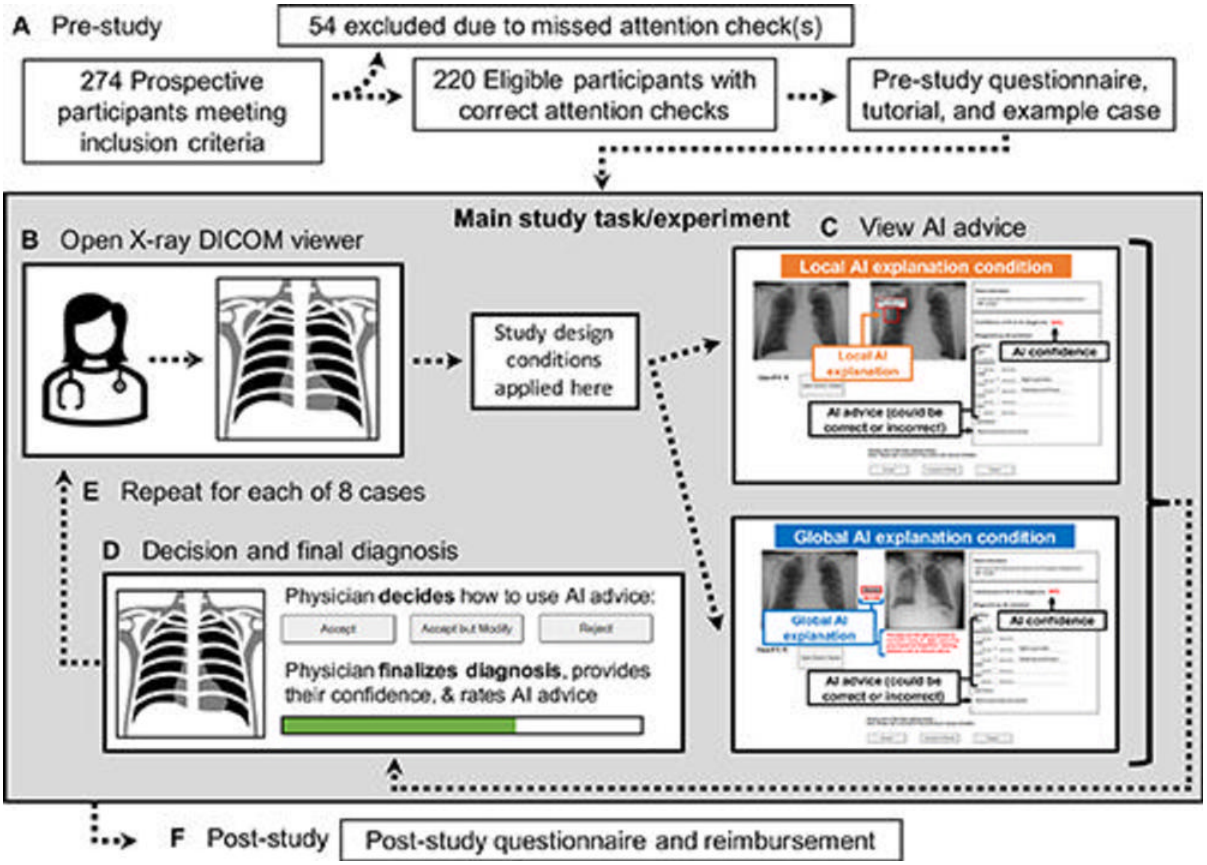


**Figure 1.** Study flow diagram. **(A)** Prestudy steps included eligibility screening and consent acquisition, followed by a prestudy questionnaire, tutorial, and example case before the main study task. **(B)** To begin the main study task, the participating physicians first viewed the radiograph case in a Digital Imaging and Communications in Medicine (DICOM) viewer without AI advice. **(C)** Once ready, participants viewed the simulated AI advice, including AI explanation and reported AI confidence level, with the design conditions applied. **(D)** The participating physicians then decided whether and how to use the AI advice, finalized their diagnosis, rated their confidence in their diagnosis, and rated the usefulness of the AI advice. **(E)** Each participant viewed eight radiographs. **(F)** Last, participants completed a poststudy questionnaire and provided their email address for reimbursement, which was not linked to the recorded study data.
High-res (TIF) version

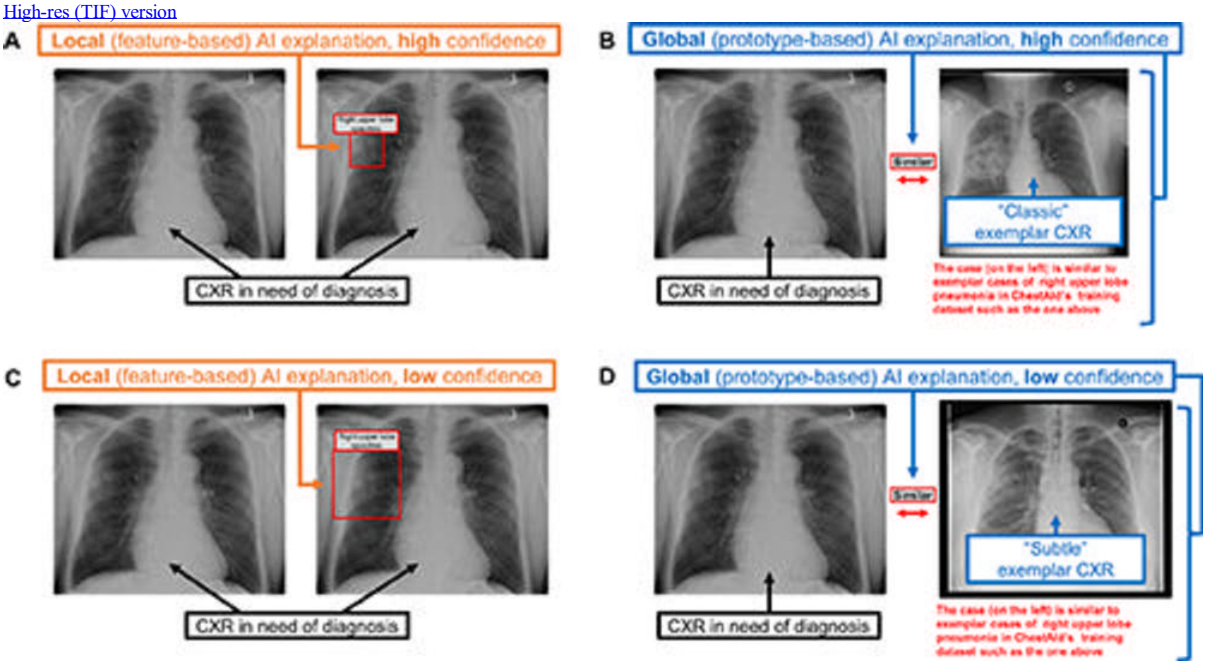


**Figure 2.** Chest radiograph (CXR) examples of **(A, C)** local (feature-based) AI explanations and **(B, D)** global (prototype-based) AI explanations from a

simulated AI tool, ChestAId, presented to physicians in the study. In all examples, the correct diagnostic impression for the radiograph case in question is "right upper lobe pneumonia," and the corresponding AI advice is correct. The patient clinical information associated with this chest radiograph was "a 63-year-old male presenting to the Emergency Department with cough." To better simulate a realistic AI system, explanation specificity was changed according to high (ie, 80%–94%) or low (ie, 65%–79%) AI confidence level: bounding boxes in high-confidence local AI explanations (example in **A**) were more precise than those in low-confidence ones (example in **C**); high-confidence global AI explanations (example in **B**) had more classic exemplar images than low-confidence ones (example in **D**), for which the exemplar images were more subtle.
High-res (TIF) version

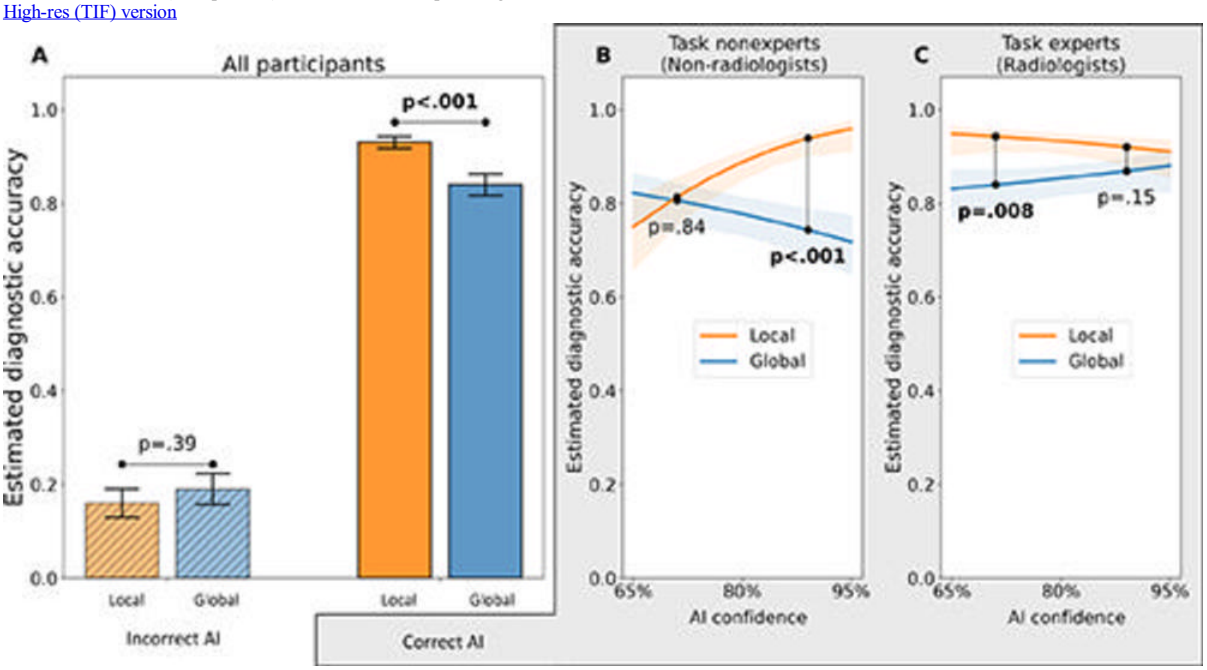


**Figure 3.** Main results for the diagnostic accuracy outcome: interaction effects among experimental variables for the outcome of marginal mean estimated diagnostic accuracy. **(A)** Interaction plot for explanation type × AI advice correctness from the generalized linear mixed-effects model (with logit link function) demonstrates that the impact of AI advice correctness on physician diagnostic accuracy depends on the type of explanation provided by the simulated AI tool. In particular, local AI explanations yielded higher diagnostic accuracy than global explanations did when AI advice was correct, whereas there was no evidence of an effect of explanation type on diagnostic accuracy when the AI advice was incorrect. **(B, C)** Interaction plots show the results for the three-way interaction of explanation type × AI advice confidence level × physician task expertise among the subset of the data corresponding to the correct AI advice condition (75% of the total data). **(B)** For task nonexperts given correct AI advice, local explanations yielded higher physician diagnostic accuracy than global explanations when AI confidence level was high, but there was no evidence of a difference when AI confidence level was low. **(C)** For task experts given correct AI advice, local explanations yielded higher diagnostic accuracy than global explanations when AI confidence level was low, but there was no evidence of such an effect when AI confidence level was high. Error bars and shaded regions represent standard errors.
High-res (TIF) version

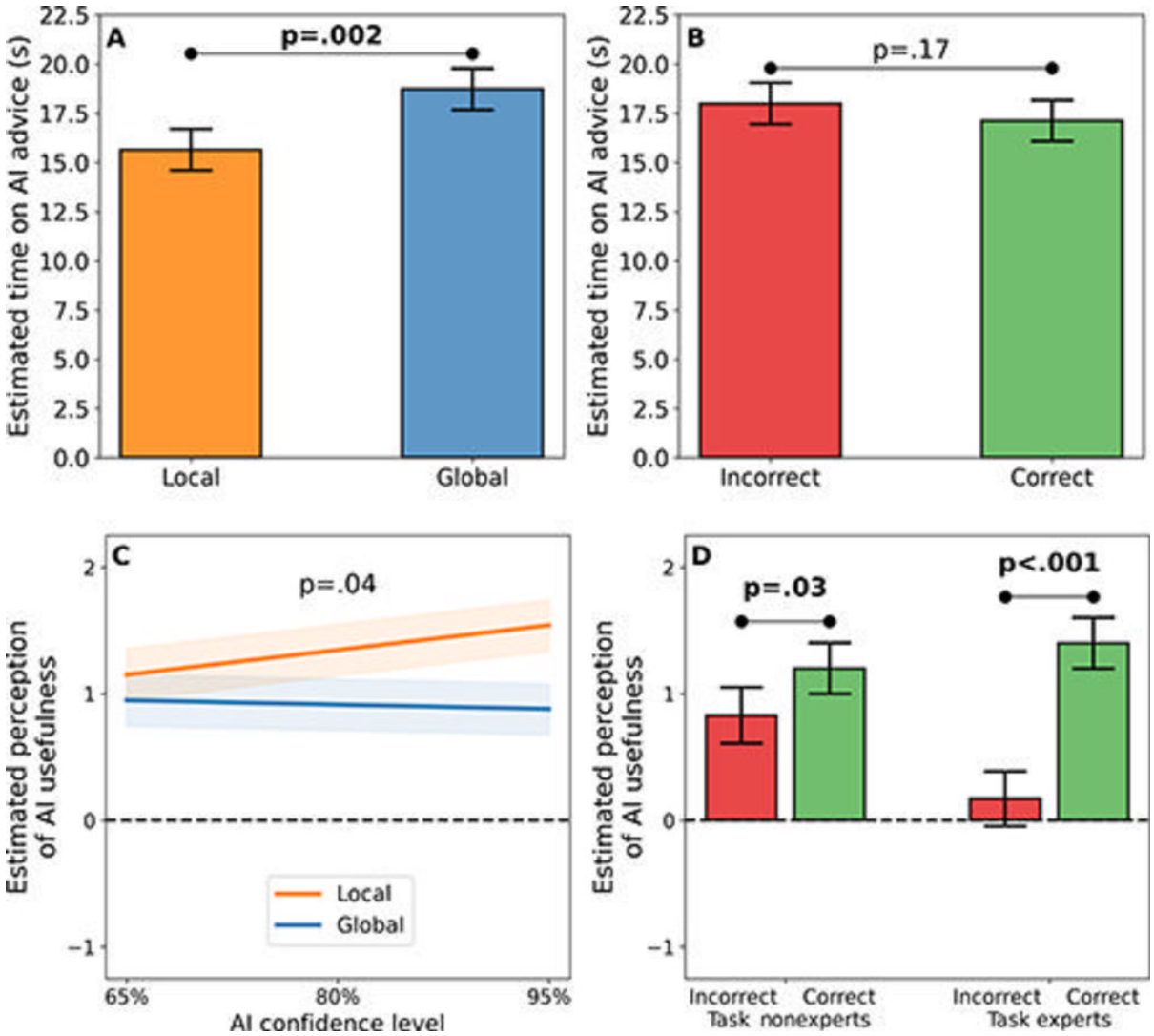


**Figure 4.** Main results for **(A, B)** the outcome of time spent viewing advice from a simulated AI tool (diagnostic efficiency) and **(C, D)** the outcome of physician perception of AI advice usefulness. **(A, B)** In these graphs, the y-axis shows the marginal mean estimated time (in seconds) spent viewing AI advice—where less time spent on AI advice indicates greater efficiency—from a log-linear mixed-effects regression model based on (A) AI explanation types and (B) AI advice correctness conditions. Local explanations yielded a more efficient decision-making process for physicians than global explanations did, but there was no evidence of a significant impact of AI advice correctness on diagnostic efficiency. (C, D) In these graphs, the y-axis shows the marginal mean estimated physician perception of AI advice usefulness (with −4 indicating the least useful, 0 indicating neutral, and 4 indicating the most useful) from a linear mixed-effects regression model. (C) The interaction effect of AI explanation type and AI confidence level on physician perception was not significant after Holm-Sidak adjustment for multiple comparisons. (D) The impact of AI advice correctness on physicians' perception of advice usefulness was greater for task experts than for task nonexperts. In particular, task experts tended to perceive a large difference in usefulness between correct and incorrect AI advice, whereas task nonexperts tended to perceive a smaller difference in usefulness between correct and incorrect AI advice.
High-res (TIF) version

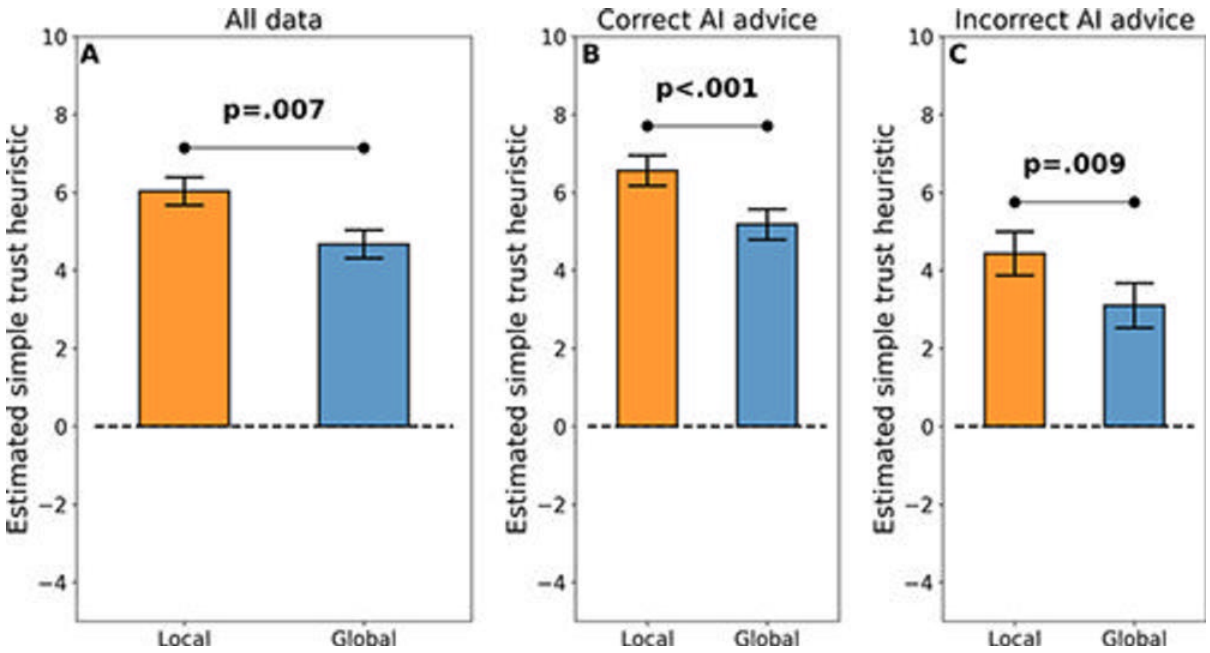


**Figure 5.** Main results for simple trust outcome. In all graphs, the y-axis shows the marginal mean estimated simple trust metric from a linear mixed-effects regression model. The simple trust metric can be understood as the speed of alignment with or divergence from advice from an AI tool, or roughly as "reliance without verification." **(A)** Local explanations promoted greater simple trust in simulated AI advice than global AI explanations across the full dataset. Moreover, this result held for the **(B)** subset of the data corresponding to correct AI advice, suggesting that local explanations could promote improved diagnostic accuracy and diagnostic efficiency when AI advice is correct. Most surprisingly, this result also held for the **(C)** subset of the data corresponding to incorrect AI advice, suggesting a potential pitfall of local explanations—that they may promote undue trust.
High-res (TIF) version

Resources:

Study abstract