Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI

Drew Prinster, MS • Amama Mahmood, MS* • Suchi Saria, PhD • Jean Jeudy, MD • Cheng Ting Lin, MD • Paul H. Yi, MD** • Chien-Ming Huang, PhD***

From the Department of Computer Science, Johns Hopkins University, 3400 N Charles St, Baltimore, MD 21218 (D.P., A.M., S.S., C.M.H.); Bayesian Health, New York, NY (S.S.); Department of Diagnostic Radiology, University of Maryland School of Medicine, Baltimore, Md (J.J., P.H.Y.); Department of Radiology, St Jude Children's Research Hospital, Memphis, Tenn (P.H.Y.); and Department of Radiology, Johns Hopkins University School of Medicine, Baltimore, Md (C.T.L.). Received December 11, 2023; revision requested January 17, 2024; final revision received September 18; accepted September 24. **Address correspondence to** C.M.H. (email: *chienming.huang@jhu.edu*).

Supported by the National Science Foundation (grant 1840088).

* D.P. and A.M. contributed equally to this work.

** P.H.Y. and C.M.H. are co–senior authors.

Conflicts of interest are listed at the end of this article.

Radiology 2024; 313(2):e233261 • https://doi.org/10.1148/radiol.233261 • Content codes: CH AI

Background: It is unclear whether artificial intelligence (AI) explanations help or hurt radiologists and other physicians in AI-assisted radiologic diagnostic decision-making.

Purpose: To test whether the type of AI explanation and the correctness and confidence level of AI advice impact physician diagnostic performance, perception of AI advice usefulness, and trust in AI advice for chest radiograph diagnosis.

Materials and Methods: A multicenter, prospective randomized study was conducted from April 2022 to September 2022. Two types of AI explanations prevalent in medical imaging—local (feature-based) explanations and global (prototype-based) explanations—were a between-participant factor, while AI correctness and confidence were within-participant factors. Radiologists (task experts) and internal or emergency medicine physicians (task nonexperts) received a chest radiograph to read; then, simulated AI advice was presented. Generalized linear mixed-effects models were used to analyze the effects of the experimental variables on diagnostic accuracy, efficiency, physician perception of AI usefulness, and "simple trust" (ie, speed of alignment with or divergence from AI advice); the control variables included knowledge of AI, demographic characteristics, and task expertise. Holm-Sidak corrections were used to adjust for multiple comparisons.

Results: Data from 220 physicians (median age, 30 years [IQR, 28–32.75 years]; 146 male participants) were analyzed. Compared with global AI explanations, local AI explanations yielded better physician diagnostic accuracy when the AI advice was correct ($\beta = 0.86$; P value adjusted for multiple comparisons [P_{adj}] < .001) and increased diagnostic efficiency overall by reducing the time spent considering AI advice ($\beta = -0.19$; $P_{adj} = .01$). While there were interaction effects of explanation type, AI confidence level, and physician task expertise on diagnostic accuracy ($\beta = -1.05$; $P_{adj} = .04$), there was no evidence that AI explanation type or AI confidence level significantly affected subjective measures (physician diagnostic confidence and perception of AI usefulness). Finally, radiologists and nonradiologists placed greater simple trust in local AI explanations than in global explanations, regardless of the correctness of the AI advice ($\beta = 1.32$; $P_{adj} = .048$).

Conclusion: The type of AI explanation impacted physician diagnostic performance and trust in AI, even when physicians themselves were not aware of such effects.

Supplemental material is available for this article.

The development of artificial intelligence (AI) diagnostic systems for both general health care and radiology has progressed rapidly in recent years, with systems offering the potential to improve patient care by assisting overburdened providers. As of 2022, 190 radiologic AI software programs had been approved by the U.S. Food and Drug Administration, with the rate of approvals increasing yearly (1). Nonetheless, a chasm has emerged between proof of concept and actual integration of AI into clinical practice (2,3). To bridge this gap, fostering appropriate trust in AI advice is paramount (4–6). Importantly, AI systems with high accuracy have demonstrated their ability to improve clinician diagnostic performance and patient outcomes in prospective, real-world settings (7–10); however, incorrect AI advice can decrease diagnostic performance (11), which rightfully contributes to delayed translational implementation.

With clinicians calling for AI tools to be transparent and interpretable (12–15), there are two broad categories of explanations that AI tools for medical imaging can provide (14,16,17): local explanations, which explain why a specific prediction was made based on a particular input (eg, highlighting informative image features on a given radiograph), and global explanations, which explain how the AI tool functions in general (eg, describing that the decision criteria of the AI tool are based on comparisons to prototypical images of each diagnostic class) (14,18–20). Additionally, clinicians often value knowing the confidence or uncertainty of the AI output for determining whether to use AI advice (12,21). Nevertheless, clinicians and AI developers disagree about the usefulness of the two main AI explanation types in health care applications (6,22). In particular, few studies have evaluated the interpretability of AI

Abbreviations

AI = artificial intelligence, DICOM = Digital Imaging and Communications in Medicine, P_{adj} = P value adjusted for multiple comparisons

Summary

In this multisite prospective study of simulated artificial intelligence (AI)–assisted chest radiograph diagnosis involving 220 physicians, AI explanation type (local vs global) differentially impacted physician diagnostic performance and trust in AI advice.

Key Results

- In a multisite prospective study of 220 physicians conducting artificial intelligence (AI)–assisted chest radiograph diagnosis, local (feature-based) AI explanations yielded better diagnostic accuracy than global (prototype-based) explanations when AI advice was correct ($\beta = 0.86$; $P_{adj} < .001$).
- Local explanations required less time to review than global explanations ($\beta = -0.19$; $P_{adj} = .01$).
- Physicians placed greater "reliance without verification" in local explanations than global explanations regardless of AI advice correctness ($\beta = 1.32$; $P_{adj} = .048$), thus potentially reducing "underreliance" on correct AI advice but increasing overreliance on incorrect advice.

explanations in medical imaging or other health care applications with human participants (23).

This multisite prospective study ($n = 220$ physician participants) aimed to address this gap by testing the hypothesis that AI advice accuracy (correct or incorrect), AI explanation type (local or global), and AI confidence level impact physician diagnostic performance (diagnostic accuracy and efficiency), physician confidence in their diagnosis, and physician perception of AI usefulness in an AI-assisted chest radiograph diagnostic task (Appendix S1). Physician task expertise, which affects physician-AI interaction dynamics (5,24), was accounted for by including both radiologists (task experts) and nonradiologist physicians (task nonexperts) as physician participants. Finally, we propose a potential "simple trust" mechanism that could underlie the main results of the study, as well as a measure for this mechanism.

Materials and Methods

This prospective experimental study was preregistered with Open Science Framework Registries *(https://osf.io/tcqfk)* and approved by the Johns Hopkins Homewood Institutional Review Board (approval number 00012741). All physician participants provided written informed consent before beginning the study (Appendix S3).

Study Participants

To be considered for eligibility, participants were required to be physicians (residents, fellows, or attending physicians) practicing in the United States who were either experts (radiologists) or nonexperts (internal or emergency medicine physicians) in chest radiograph diagnosis. Recruitment was conducted through U.S. medical school mailing lists and direct emails to colleagues of radiologist team members, and the study was conducted from April 2022 to September 2022. Data from any participant who failed an attention check in either the prestudy or poststudy questionnaires were excluded from analyses (Fig 1). Participants were compensated with a $10 gift card. Self-reported participant demographic characteristics were collected for transparency and reproducibility (ie, to facilitate comparison with other studies that may differ in study population). See Appendix S2 for further details.

Radiograph Selection and AI Advice

The chest radiograph diagnosis task consisted of eight clinical vignettes, each using frontal (anteroposterior or posteroanterior) and, if available, corresponding lateral chest radiograph projections obtained from Beth Israel Deaconess Hospital (Boston, Massachusetts) via the open-source MIMIC-CXR Database *(https://physionet.org/content/mimic-cxr/2.0.0/)*. A panel of radiologists (5) previously selected the cases and generated the associated AI advice, to produce a set that well simulated real clinical practice and allowed known challenges in radiograph assessment to be evaluated. Additionally, three board-certified and fellowship-trained radiologists (J.J., C.T.L., and P.H.Y., with 18, 9, and 2 years of postresidency experience, respectively) reviewed the suitability of the cases for the present study (Appendixes S4–S6).

Within-Participant Factors: AI Correctness and AI Confidence Level

The correctness and confidence of the AI advice were randomly varied across the eight cases that each participant reviewed. Specifically, the participants were randomly assigned six cases paired with AI-presented findings and impressions that were correct and two cases paired with incorrect advice to simulate a state-of-the-art AI diagnostic tool, which would be correct more often than it would be incorrect. The AI confidence level for each suggestion was a randomly selected integer percentage in the range of 65%–94% (Appendix S7).

Between-Participant Factor: AI Explanation Type

Between participants, whether the cases were paired with local (feature-based) explanations or global (prototype-based) explanations was randomly varied. Local explanations were presented as annotated bounding boxes on the chest radiograph that identified anomalous or important regions informing the AI tool's diagnostic advice (eg, Fig 2A). Global explanations were presented as a visual comparison between the chest radiograph in question and a "prototypical" example radiograph for the diagnosis provided by the AI tool, with text explaining that the AI tool identified the case image as similar to this exemplar image from the AI training dataset (eg, Fig 2B). The team of three expert radiologists generated both the local and global AI explanations used in this study to mimic a realistic, high-performing AI system as closely as possible. For instance, when AI confidence level was higher, local explanation bounding boxes were more precise (eg, Fig 2A vs Fig 2C), and global explanation exemplar images were more "classic" rather than "subtle" (eg, Fig 2B vs Fig 2D). See Appendix S8 for further details.

Study Procedure

This study was conducted online through a website interface. Before the main task, the website displayed standard informed consent information, a questionnaire about the participant's opinions on AI in radiology, a tutorial, and a practice case (Fig 1).

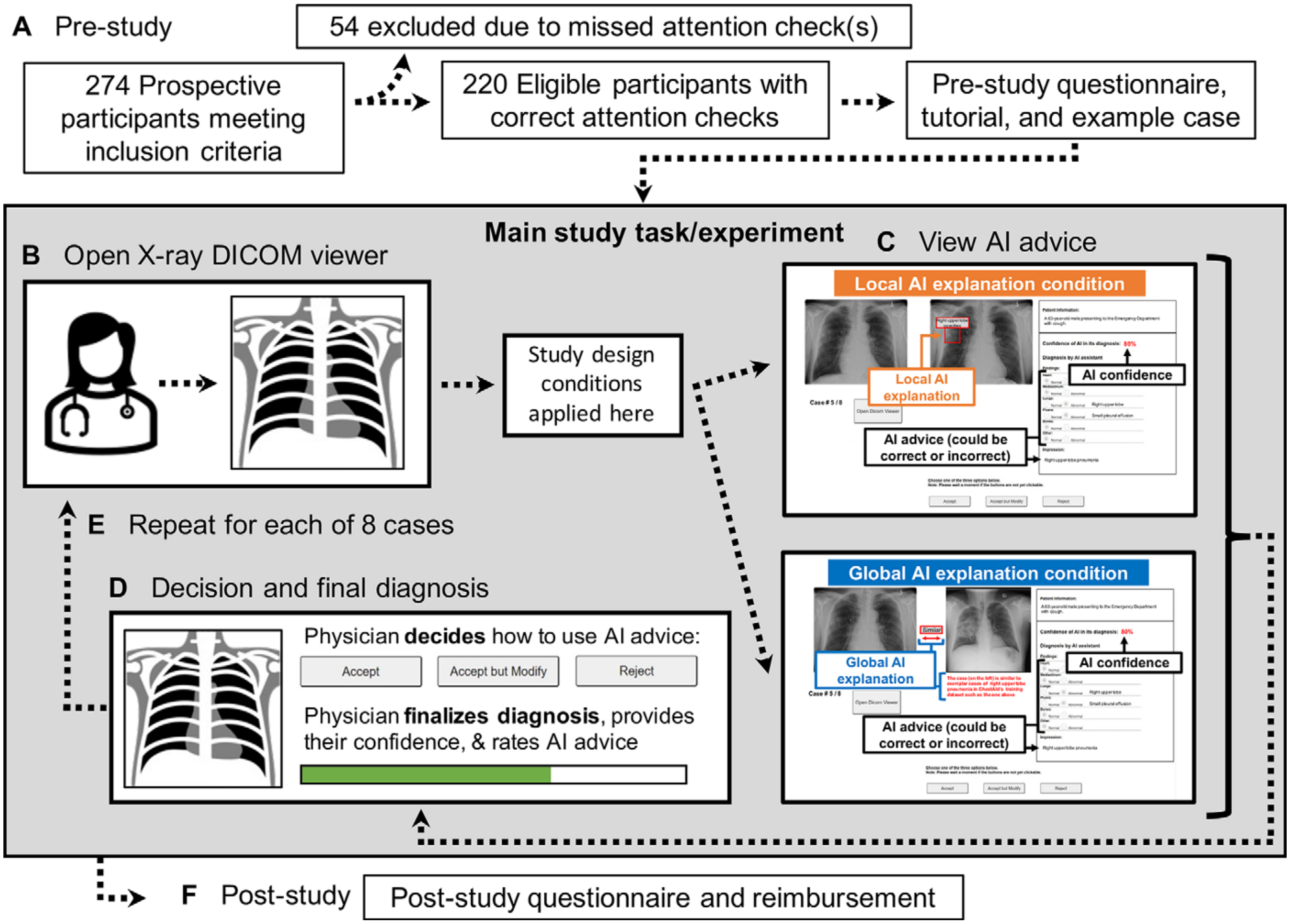
A Pre-study
54 excluded due to missed attention check(s)
274 Prospective participants meeting inclusion criteria
220 Eligible participants with correct attention checks
Pre-study questionnaire, tutorial, and example case
Main study task/experiment
B Open X-ray DICOM viewer
Study design conditions applied here
C View AI advice
Local AI explanation condition
Right upper lobe opacities
Local AI explanation
AI confidence
AI advice (could be correct or incorrect)
Case # 5 / 8
Open Dicom Viewer
Patient Information:
A 63-year-old male presenting to the Emergency Department with cough.
Confidence of AI in its diagnosis: 80%
Diagnosis by AI assistant
Findings:
Heart:
Mediastinum:
Lungs:
Pleura:
Bones:
Other:
Normal
Abnormal
Right upper lobe
Small pleural effusion
Impression:
Right upper lobe pneumonia
Choose one of the three options below.
Note: Please wait a moment if the buttons are not yet clickable.
Accept
Accept but Modify
Reject
Global AI explanation condition
Similar
Global AI explanation
The case (on the left) is similar to exemplar cases of right upper lobe pneumonia in ChestAId's training dataset such as the one above
E Repeat for each of 8 cases
D Decision and final diagnosis
Physician decides how to use AI advice:
Accept
Accept but Modify
Reject
Physician finalizes diagnosis, provides their confidence, & rates AI advice
F Post-study
Post-study questionnaire and reimbursement

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 artificial intelligence (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.

The primary study task was to evaluate eight chest radiograph cases with suggestions from a purported AI assistant called ChestAId, which was stated to be a deep learning–based AI tool with diagnostic performance comparable to that of experts in the field. For each case, the participant was first presented with a short note on the clinical history of the patient and then was required to open the radiograph in a fully functional browser-embedded Digital Imaging and Communications in Medicine (DICOM) viewer. Once the participant indicated that they were ready to proceed, the AI advice was presented next to the chest radiograph, along with an "open DICOM viewer" button to allow the participant to return to the radiograph viewer at will. The AI suggestions varied according to the assigned advice correctness, AI explanation type, and AI confidence level for the case. The participant could make one of three responses to the AI advice: accept (ie, use the AI suggestion as is, without modification), modify (ie, keep the prefilled AI suggestion but modify it freely), or reject (ie, completely discard the AI suggestion and complete the diagnosis independently). After selecting an option, the participant finalized the diagnosis and reported their own confidence level (0%–100%) in their findings and impressions. Finally, the participant provided two Likert scale ratings of the usefulness of the AI advice for each case. (Further details in Appendix S3.)

Statistical Analysis

Recruitment was stopped once the number of included participants reached 220, a sample size comparable to that of previous physician-AI interaction studies with a similar experimental design (5,24). One author (D.P.) fit generalized linear mixed-effects models for each outcome variable using the *lme4* package (25) in R (version 4.2.0) (26). Various model structures were used, based on their appropriateness given the distributions of the dependent variables (see Results for descriptions of outcome metrics and Appendix S10 for analysis details). All regressions included the three experimentally manipulated variables and their interactions, along with the control covariates age, gender, knowledge of AI, AI aversion, job title (trainee [resident or fellow] or attending), years in practice, task expert or nonexpert, and engagement as measured by DICOM view time (the total time in seconds that a participant spent on the radiograph viewer for a given case). Interaction effects involving any of the three experimental variables and either task expertise or the engagement measure were assessed. Holm-Sidak corrections were used to control the familywise error rate

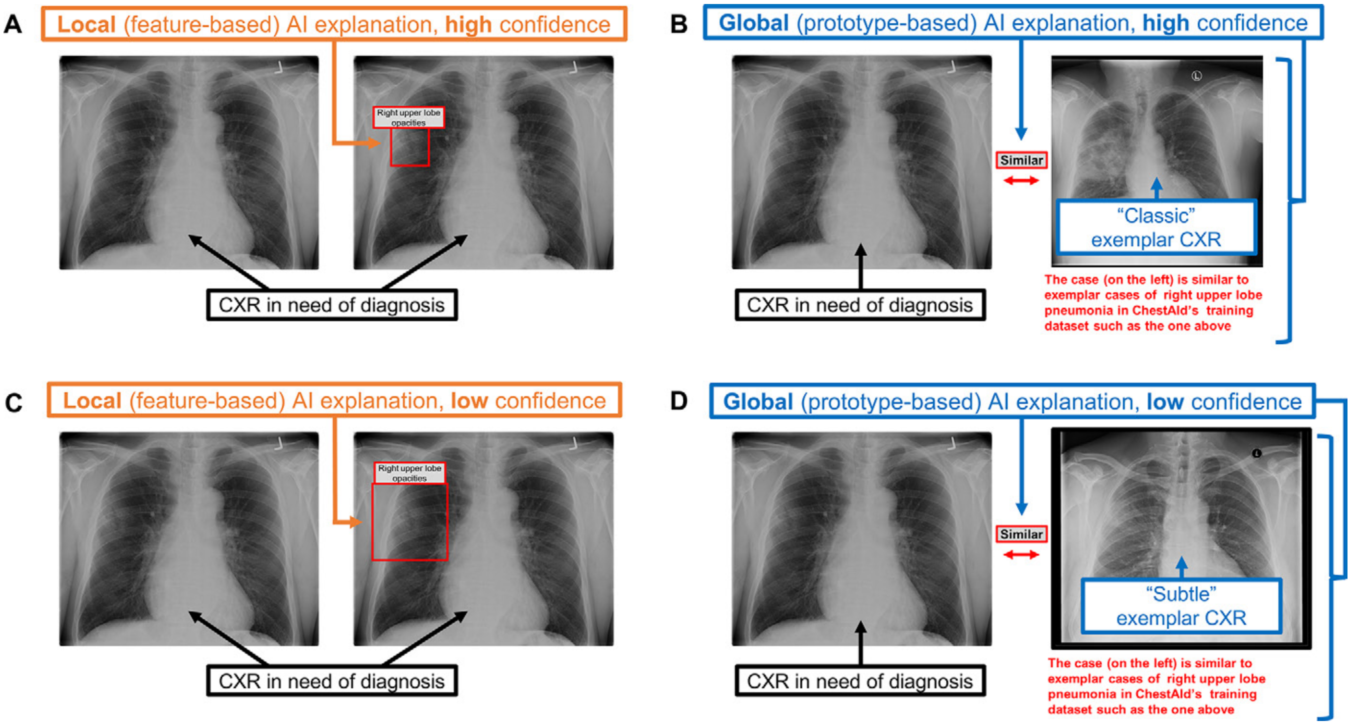
A
Local (feature-based) AI explanation, high confidence
Right upper lobe opacities
CXR in need of diagnosis
B
Global (prototype-based) AI explanation, high confidence
Similar
"Classic" exemplar CXR
CXR in need of diagnosis
The case (on the left) is similar to exemplar cases of right upper lobe pneumonia in ChestAId's training dataset such as the one above
C
Local (feature-based) AI explanation, low confidence
Right upper lobe opacities
CXR in need of diagnosis
D
Global (prototype-based) AI explanation, low confidence
Similar
"Subtle" exemplar CXR
CXR in need of diagnosis
The case (on the left) is similar to exemplar cases of right upper lobe pneumonia in ChestAId's training dataset such as the one above

Figure 2: Chest radiograph (CXR) examples of **(A, C)** local (feature-based) artificial intelligence (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.

at a significance level α of .05 for each model and for post hoc pairwise comparisons of interaction effects (Appendix S10).

Results

The main study results are presented in this section, and the full results are provided in Appendix S11.

Participant Characteristics

Among the initial 274 participants, 54 were excluded due to missed attention checks (Fig 1A). All analyses were conducted with data from 220 physician participants (median age, 30 years [IQR, 28–32.75 years]; 146 male participants). These physicians had been in practice for a median 3 years (IQR, 2–4.25 years), and 132 were radiologists (task experts); detailed self-reported social and professional demographic characteristics are presented in Table 1.

Outcome Metrics

Table 2 presents the quantitative metrics used to evaluate physicians' diagnostic performance (diagnostic accuracy and efficiency), their confidence in their diagnosis, the perceived usefulness of AI suggestions, and simple trust in AI advice (see Appendix S9 for details on the simple trust outcome). Table 3 provides a breakdown of case-specific accuracy and time spent on AI advice. When AI advice was correct, physicians' diagnostic accuracy was 92.8% ± 0.62 (standard error) with local explanations and 85.3% ± 0.85 with global explanations; with incorrect AI advice, physician diagnostic accuracy was 23.6% ± 1.02 with local explanations and 26.1% ± 1.06 with global explanations.

Impact of AI Advice Correctness and AI Explanation Type on Diagnostic Accuracy

The impact of AI advice correctness on physician diagnostic accuracy depended on the AI explanation type (interaction coefficient β = 1.09; P = .001 [P value adjusted for multiple comparisons {P_{adj}} = 0.01]) (Fig 3A). In other words, the benefits of correct AI advice (or the detriments of incorrect AI advice) were affected by the type of AI explanation associated with the advice. In particular, in post hoc pairwise comparisons, local explanations yielded better diagnostic accuracy than global explanations did when AI advice was correct (β = 0.86; P < .001 [P_{adj} < .001]) (Fig 3A). There was no evidence of a differential impact of explanation type when AI advice was incorrect (β = −0.23; P = .39 [P_{adj} = .39]) (Fig 3A), though this condition may have been underpowered (see Appendixes S7, S10, and S11).

Moreover, when AI advice was correct, the benefit of local explanations over global explanations depended on the interaction between AI confidence level and physician task expertise (three-way interaction of explanation type, AI confidence level, and task expertise: β = −1.05; P = .01 [P_{adj} = .04]) (Fig 3B, 3C). In particular, in post hoc pairwise comparisons for correct AI advice, task nonexperts (nonradiologists) benefitted from local explanations only when the AI tool indicated high confidence (β = 1.62; P < .001 [P_{adj} = .002]) (Fig 3B), whereas for task experts (radiologists), a benefit of local explanations was observed only when the AI tool indicated low confidence (β = 1.12; P = .008 [P_{adj} = .03]) (Fig 3C).

Impact of AI Explanation Type on Physician Efficiency, Perception of AI Usefulness, and Confidence in Diagnosis

Local explanations led to higher efficiency than global explanations did because physicians spent less time considering AI advice with local explanations ($\beta = -0.19$; $P = .002$ [$P_{adj} = .01$]) (Fig 4A). No evidence of an effect of advice correctness on diagnostic efficiency was found ($\beta = -0.06$; $P = .17$ [$P_{adj} = .53$]) (Fig 4B).

Table 1: Self-reported Social and Professional Demographic Characteristics of the Study Participants

Characteristic	Value
Age (y)	
Mean[*]	32.00 ± 6.82
Median[†]	30 (28–32.75)
Range	25–68
Gender	
Female	69
Male	146
Nonbinary	1
Other	4
Race or ethnicity	
Asian	65
Black or African American	11
White	104
Hispanic, Latino, or Spanish origin	14
Native American or Alaskan Native	1
Native Hawaiian or other Pacific Islander	0
Prefer not to say	31
Designation	
Resident	166
Fellow	14
Attending	40
Expertise	
Expert (radiologist)	132
Nonexpert (internal or emergency medicine physician)	88
Years in practice[†]	3 (2.0–4.25)

Note.—Categorical data are presented as numbers of participants.

[*] Continuous data presented as means ± SDs.

[†] Continuous data presented as medians, with IQRs in parentheses.

Regarding physician perception of AI advice usefulness, there was no evidence of an effect of AI explanation type ($\beta = 0.35$; $P = .07$ [$P_{adj} = .37$]) or AI confidence level ($\beta = -0.16$; $P = .22$ [$P_{adj} = .53$]); moreover, the interaction effect of AI explanation type and AI confidence level on perceived usefulness was ultimately not significant after Holm-Sidak adjustment for multiple comparisons ($\beta = 0.40$; $P = .04$ [$P_{adj} = .24$]) (Fig 4C). However, there was an interaction effect of physician task expertise and AI advice correctness on perceived usefulness ($\beta = 0.84$; $P < .001$ [$P_{adj} = .002$]). Notably, post hoc pairwise comparisons revealed that the impact of AI advice correctness on physician perception of AI advice usefulness was greater for task experts than for task nonexperts (Fig 4D).

Physicians' confidence in their final diagnosis was higher with correct AI advice than with incorrect AI advice ($\beta = 0.08$; $P = .001$ [$P_{adj} = .01$]), but there was no significant effect of AI explanation type ($\beta = 0.02$; $P = .32$ [$P_{adj} = .90$]) or AI confidence level on physician confidence ($\beta = -0.004$; $P = .82$ [$P_{adj} > 0.99$]).

Underlying "Simple Trust" Mechanism: Effects and Caveats

The last key result of this study suggests that the apparent benefits of local explanations for improved diagnostic accuracy and efficiency could be attributed to physicians' placing greater simple trust in local explanations than in global explanations (Fig 5). The proposed metric for simple trust (Table 2, Appendix S9) is the speed at which a physician aligns their final diagnosis with (or diverges their final diagnosis from) AI advice and can also be understood as "reliance without verification." Overall, physicians were more likely to align their decision with AI advice and undergo a shorter period of consideration when local explanations were given than when global explanations were given ($\beta = 1.32$; $P = .007$ [$P_{adj} = .048$]) (Fig 5A). When this analysis was repeated separately for correct and incorrect AI advice, local AI explanations increased simple trust in AI advice regardless of AI advice correctness. The fact that this result held when the analysis was limited to correct AI advice ($\beta = 1.37$; $P < .001$

Table 2: Definitions of Quantitative Outcomes

Outcome	Definition
Behavioral outcomes	
Diagnostic accuracy	Proportion of cases for which a participant provided the correct impression
Diagnostic efficiency	Time (seconds) spent considering the AI advice before proceeding (with less time indicating more efficiency)
Simple trust	Percent alignment with (or divergence from) AI advice per second, with large positive values indicating quick alignment, large negative values indicating quick divergence, and values close to zero indicating slower alignment or divergence (see Appendix S9 for details)
Subjective outcomes	
Perceived usefulness of AI advice	Self-reported scale ranging from −4 to 4 computed by adding up the values from two Likert scale questions (range, −2 to 2, with the labels −2 = strongly disagree, 0 = neutral, 2 = strongly agree); the two questions were "The AI's recommendation was useful" and "I would consult this AI assistant for future diagnosis tasks"
Physician confidence in impression	Physicians were asked to rate their confidence in their impression on a scale of 0%–100% when they submitted their diagnosis
Physician confidence in findings	Physicians were asked to rate their confidence in their findings on a scale of 0%–100% when they submitted their diagnosis

Note.—AI = artificial intelligence.

Table 3: Physician Diagnostic Accuracy and Time Spent Viewing AI Advice in Judging Eight Radiograph Cases with Different AI Correctness and Explanation Type Conditions

Case No. and AI Advice Correctness*	AI Explanation Type	Physician Diagnostic Accuracy[†]	Mean Time (sec) Spent on AI Advice[‡]
Case 1			
Correct (normal)	Global	85/94 (90 ± 0.71)	13.25 ± 0.27
	Local	66/70 (94 ± 0.56)	17.32 ± 0.66
Incorrect (lingular pneumonia)	Global	11/28 (39 ± 1.19)	30.23 ± 0.62
	Local	6/28 (21 ± 1.00)	16.49 ± 0.33
Case 2			
Correct (right sternoclavicular dislocation)	Global	85/90 (94 ± 0.55)	18.45 ± 0.34
	Local	75/75 (100 ± 0.00)	13.85 ± 0.21
Incorrect (right pleural plaque)	Global	9/30 (30 ± 1.12)	26.13 ± 0.32
	Local	3/23 (13 ± 0.83)	19.99 ± 0.33
Case 3			
Correct (hiatus hernia)	Global	78/97 (80 ± 0.96)	31.31 ± 0.62
	Local	71/73 (97 ± 0.39)	19.88 ± 0.25
Incorrect (normal)	Global	5/25 (20 ± 0.98)	22.70 ± 0.26
	Local	5/23 (22 ± 1.01)	20.58 ± 0.33
Case 4			
Correct (right pneumothorax)	Global	86/89 (97 ± 0.44)	26.18 ± 0.39
	Local	67/69 (97 ± 0.41)	20.14 ± 0.35
Incorrect (right lower lobe pneumonia)	Global	8/31 (26 ± 1.07)	21.38 ± 0.33
	Local	7/26 (27 ± 1.09)	25.77 ± 0.38
Case 5			
Correct (right upper lobe pneumonia)	Global	93/97 (96 ± 0.48)	17.18 ± 0.22
	Local	72/77 (94 ± 0.60)	14.51 ± 0.29
Incorrect (pulmonary venous congestion)	Global	14/24 (58 ± 1.21)	16.03 ± 0.26
	Local	10/21 (48 ± 1.23)	23.48 ± 0.54
Case 6			
Correct (rib fracture)	Global	42/91 (46 ± 1.20)	27.47 ± 0.54
	Local	45/70 (64 ± 1.16)	20.60 ± 0.30
Incorrect (pulmonary nodule)	Global	2/31 (6.5 ± 0.60)	23.87 ± 0.75
	Local	6/27 (22 ± 1.02)	17.71 ± 0.55
Case 7			
Correct (pulmonary edema)	Global	78/87 (90 ± 0.74)	22.39 ± 0.54
	Local	71/72 (99 ± 0.28)	25.51 ± 1.15
Incorrect (left lower lobe pneumonia)	Global	6/30 (20 ± 0.98)	22.49 ± 0.37
	Local	4/24 (17 ± 0.91)	17.51 ± 0.26
Case 8			
Correct (left pneumothorax)	Global	70/78 (90 ± 0.73)	25.52 ± 0.47
	Local	74/77 (96 ± 0.47)	20.38 ± 0.38
Incorrect (left lower lobe pneumonia)	Global	7/39 (18 ± 0.93)	23.94 ± 0.28
	Local	4/19 (21 ± 1.01)	14.59 ± 0.26
All cases			
Correct	Global	617/723 (85.3 ± 0.85)	22.65 ± 0.46
	Local	541/583 (92.8 ± 0.62)	18.97 ± 0.54
Incorrect	Global	62/238 (26.1 ± 1.06)	23.50 ± 0.44
	Local	45/191 (23.6 ± 1.02)	19.55 ± 0.39

Note.—Data from 220 physician participants (total of 1735 measurements). AI = artificial intelligence.

* The correct and incorrect diagnosis provided by the simulated AI tool for each case is given in parentheses.

[†] Data presented as numbers of accurate cases out of the total number of cases, with percentages ± standard errors in parentheses.

[‡] Data presented as means ± standard errors.

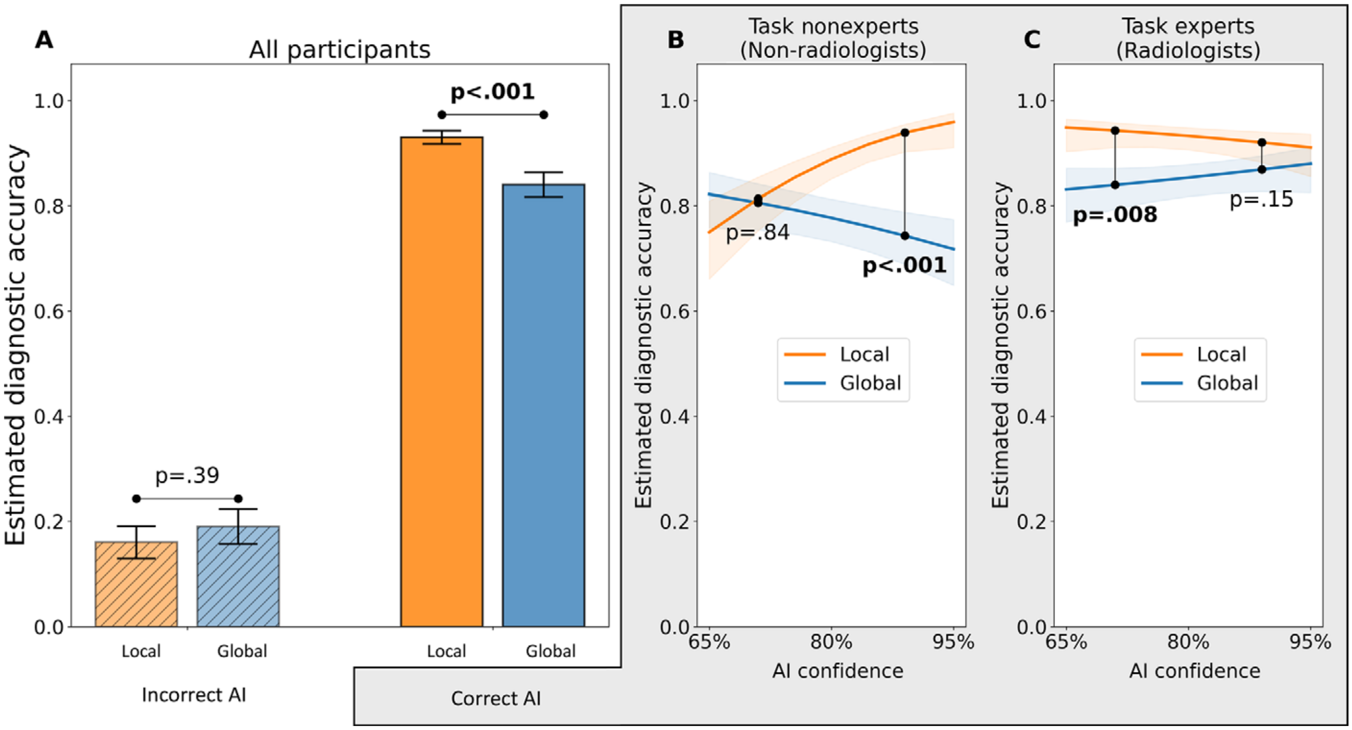

Figure 3: Main results for the diagnostic accuracy outcome: interaction effects among experimental variables for the outcome of marginal mean estimated diagnostic accuracy. **(A)** Interaction plot for explanation type × artificial intelligence (AI) advice correctness (interaction coefficient β = 1.09; *P* = .001 [*P* value adjusted for multiple comparisons {P_{adj}} = 0.01]) 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 (β = 0.86; *P* < .001 [P_{adj} < .001]), whereas there was no evidence of an effect of explanation type on diagnostic accuracy when the AI advice was incorrect (β = −0.23; *P* = .39 [P_{adj} = .39]). **(B, C)** Interaction plots show the results for the three-way interaction of explanation type × AI advice confidence level × physician task expertise (β = −1.05; *P* = .01 [P_{adj} = .04]) 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 (β = 1.62; *P* < .001 [P_{adj} = .002]), but there was no evidence of a difference when AI confidence level was low (β = 0.07; *P* = .84 [P_{adj} = .84]). **(C)** For task experts given correct AI advice, local explanations yielded higher diagnostic accuracy than global explanations when AI confidence level was low (β = 1.12; *P* = .008 [P_{adj} = .02]), but there was no evidence of such an effect when AI confidence level was high (β = 0.56; *P* = .15 [P_{adj} = .28]). Error bars and shaded regions represent standard errors.

[P_{adj} < .001]) (Fig 5B) suggests a mechanism to explain how local explanations improve diagnostic accuracy (Fig 3). However, the fact that the result held when the analysis was limited to incorrect AI advice (β = 1.32; *P* = .009 [P_{adj} = .03]) (Fig 5C) suggests that local explanations could pose a risk of exacerbating overreliance on AI when AI advice is incorrect.

Discussion

For artificial intelligence (AI) to realize its full potential in improving clinician performance in radiologic diagnosis and consequently patient outcomes, it is essential to design AI interfaces for optimal human-machine teamwork (5,6). Our study investigated how to communicate AI insights effectively to clinicians to achieve such collaboration and found that different AI explanation methods yielded disparate collaboration outcomes. In particular, in the popular radiologic AI use case of disease diagnosis based on chest radiography, local (feature-based) explanations encouraged greater physician simple trust in simulated AI advice than global (prototype-based) explanations (β = 1.32; *P* value adjusted for multiple comparisons [P_{adj}] = .048). This influence on physician trust corresponded with improved diagnostic accuracy when the AI advice was correct (β = 0.86; P_{adj} < .001), and overall diagnostic efficiency was higher (ie, the time spent on AI advice was reduced) for local AI explanations than for global AI explanations (β = −0.19; P_{adj} = .01). Additionally, there were nuanced interactions of AI confidence level, physician task expertise, and AI explanation type (three-way interaction coefficient β = −1.05; P_{adj} = .04). Future developers and users of health care AI systems should therefore pay careful consideration to how different forms of AI explanations, along with other considerations such as AI uncertainty presentation and users' experience level, might impact reliance on AI advice and support optimal collaboration with the technology across various clinical use cases (6).

Here, local AI explanations were more effective than global explanations in reinforcing the benefits of correct AI advice. Prior research has shown that local explanations lead to better decision-making relative to a lack of explanation for AI-assisted chest radiograph diagnosis, assuming that the AI advice is correct (24). Our findings substantially extend this result by revealing how an AI explanation can have variable impact depending on the type of AI explanation provided and whether the AI advice is correct. Moreover, our results for correct AI advice further clarify that the effect of local versus global explanations may also depend on presented AI confidence level and physician task expertise. For task nonexperts, local explanations were more effective only when AI confidence level was high, whereas for task experts, local explanations were more effective only when AI confidence level was low. This task expertise–dependent impact of AI confidence level

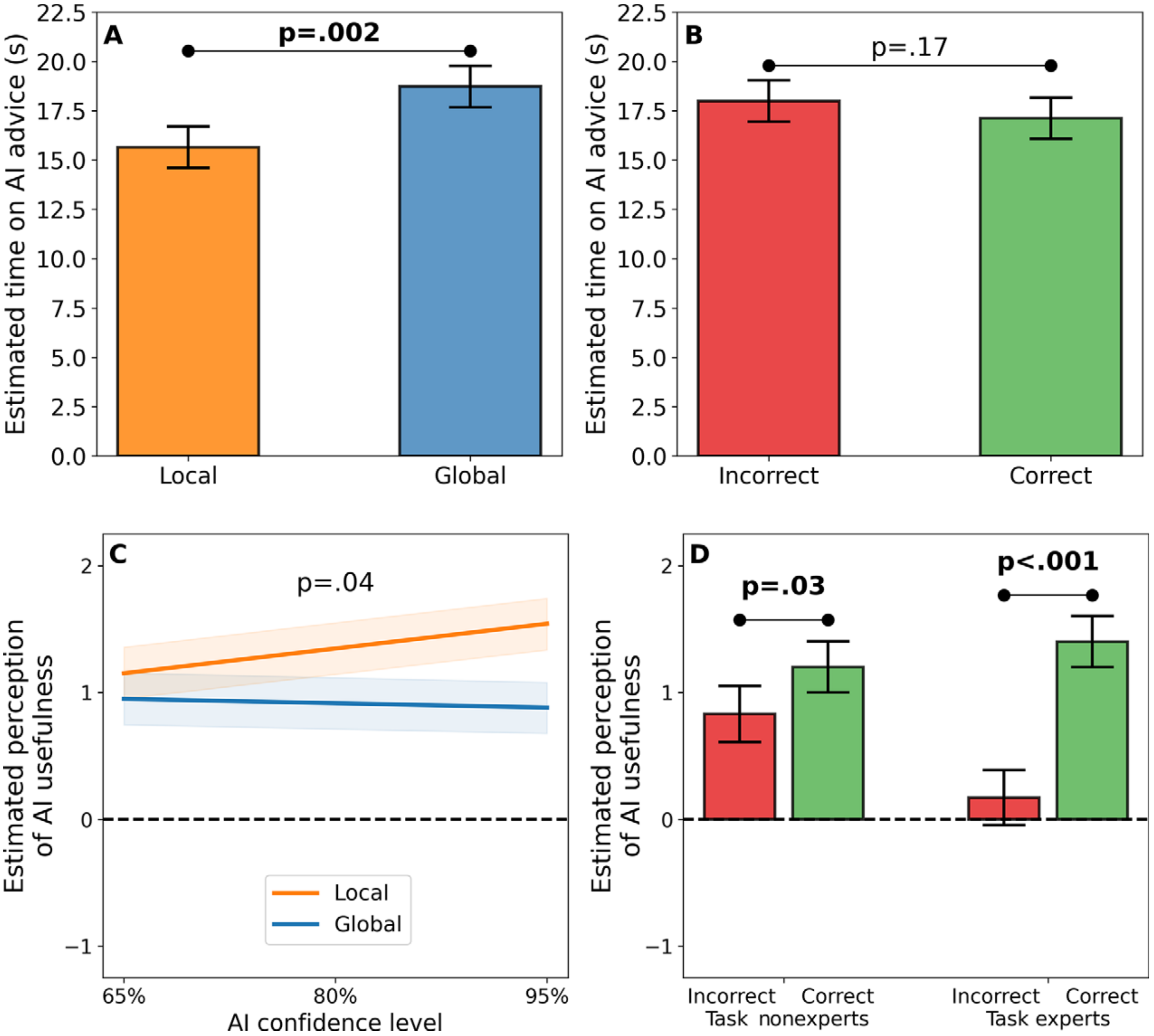

Figure 4: Main results for **(A, B)** the outcome of time spent viewing advice from a simulated artificial intelligence (AI) tool (diagnostic efficiency) and **(C, D)** the outcome of physician perception of AI advice usefulness. **(A, B)** In these graphs, the y-axis shows the marginal mean estimated 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 (β = −0.19; *P* = .002 [*P* value adjusted for multiple comparisons {P_{adj}} = 0.01]), but there was no evidence of a significant impact of AI advice correctness on diagnostic efficiency (β = −0.06; *P* = .17 [P_{adj} = .53]). **(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 (β = 0.40; *P* = .04 [P_{adj} = .24]). **(D)** The impact of AI advice correctness on physicians' perception of advice usefulness was greater for task experts than for task nonexperts (β = 0.84; *P* < .001 [P_{adj} = .002]). In particular, task experts tended to perceive a large difference in usefulness between correct and incorrect AI advice (β = 1.23; *P* < .001 [P_{adj} < .001]), whereas task nonexperts tended to perceive a smaller difference in usefulness between correct and incorrect AI advice (β = 0.39; *P* = .03 [P_{adj} = .03]).

could be attributed to task nonexperts being especially swayed by high-confidence local explanations and task experts being especially skeptical of low-confidence global explanations. When the AI advice was incorrect, however, it was inconclusive whether explanation type differentially affected diagnostic accuracy; this null result may be due to reduced statistical power in the incorrect AI advice condition.

Reducing the cognitive load of clinicians is a crucial goal that is complementary to efforts to improve diagnostic accuracy (27); this is especially true in radiology, given its high rates of burnout (28) and fatigue-linked diagnostic errors (29). Our study revealed that physicians were able to determine a diagnosis more quickly when provided with local AI explanations than with global explanations. Previous studies have variously found beneficial (30) and detrimental (31) effects of explainable AI advice on physician efficiency: Our results highlight that the specific type of explanation provided could be a key factor underlying such variability.

Correct AI advice positively affected the subjective outcomes of physician confidence in the final diagnosis and physician perception of AI advice usefulness. Moreover, this effect on perceived usefulness was greater for task experts, suggesting that they were better able to discern the difference between correct and incorrect AI advice. However, unlike previous studies (32), no significant effects of AI explanation type or AI confidence level on these subjective outcomes

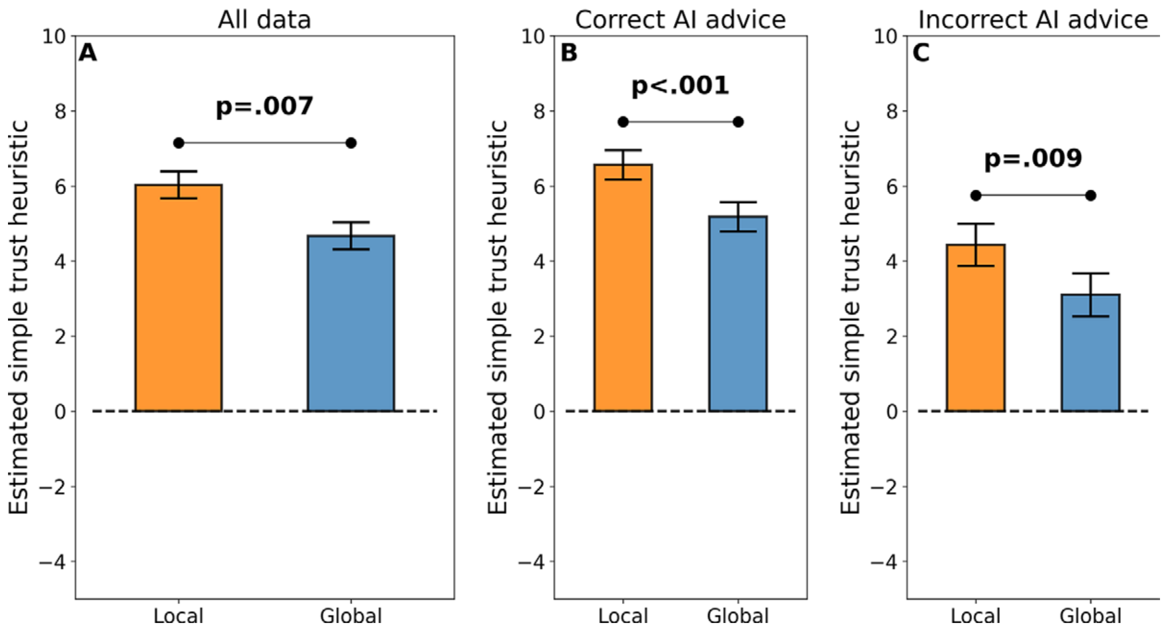

Figure 5: Main results for simple trust outcome. In all graphs, the y-axis shows the marginal mean estimated simple trust metric (Appendix S9) from a linear mixed-effects regression model. The simple trust metric can be understood as the speed of alignment with or divergence from advice from an artificial intelligence (AI) tool, or roughly as "reliance without verification." **(A)** Local explanations promoted greater simple trust in simulated AI advice than global AI explanations (β = 1.32; *P* = .007 [*P* value adjusted for multiple comparisons {P_{adj}} = 0.048]) across the full dataset. Moreover, this result held for the **(B)** subset of the data corresponding to correct AI advice (β = 1.37; *P* < .001 [P_{adj} < .001]), 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 (β = 1.32; *P* = .009 [P_{adj} = .03]), suggesting a potential pitfall of local explanations—that they may promote undue trust.

were observed. The discrepancy between the influence of AI explanation types on behavioral diagnostic outcomes (diagnostic accuracy and efficiency) and the lack of impact on subjective measures (physician perception and confidence) underscores that AI explanation types can differentially affect physicians even if the physicians themselves do not recognize these effects.

While the main results of this study demonstrate that local explanations enhance diagnostic accuracy and efficiency when AI advice is correct, the further analyses of trust calibration conducted in this study suggest both a possible mechanism underlying these benefits and a critical caveat. Local explanations foster simple trust in AI—as defined by how swiftly physicians align with AI advice (33,34)—regardless of its correctness. This means that while local explanations can reduce "underreliance" on correct AI advice, they may also lead to overreliance on incorrect AI advice, potentially increasing the risk of AI-related errors.

The limitations of our study offer opportunities for future work. Despite having a fully functional DICOM viewer, the web interface used here could not fully replicate the features and nuances of actual radiologic practice, such as voice transcription, real time pressures, and sensitivity to real patient outcomes. Additionally, presenting more chest radiograph cases to each participant could help elucidate case-specific effects, and presenting more cases with incorrect AI advice could help improve statistical power for detecting the potential role of explanation type in exacerbating AI-related errors. Exploring other forms or combinations of AI explanations and alternative representations of AI confidence could also yield valuable insights.

In simulated artificial intelligence (AI)–assisted chest radiograph diagnosis, the impact of AI explanations on physicians' diagnostic performance and trust in AI advice depended on the specific type of AI explanation provided, even when physicians themselves were not aware of such effects. Future efforts to develop AI decision support systems should thus take care to account for the intricate influences of design elements, with particular attention given to the distinct impacts of different explanation types.

Deputy Editor: Linda Moy
Scientific Editor: Shannyn Wolfe (AJE)

Author contributions: Guarantors of integrity of entire study, **D.P., A.M., P.H.Y., C.M.H.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **D.P., A.M., P.H.Y.**; clinical studies, **P.H.Y.**; experimental studies, all authors; statistical analysis, **D.P., A.M., S.S.**; and manuscript editing, all authors

Data sharing: Data generated by the authors or analyzed during the study are available at the repository of our Open Science Framework preregistration: *https://osf.io/gvbmp*.

Disclosures of conflicts of interest: D.P. No relevant relationships. **A.M.** No relevant relationships. **S.S.** Holds equity in Bayesian Health and Duality Tech; serves on the advisory boards of Fortune 500 and health technology companies; honoraria for talks and travel related to her work from biotechnology and technology companies and academic health systems; and grants from the Gordon and Betty Moore Foundation, National Science Foundation, National Institutes of Health, Defense Advanced Research Projects Agency, Food and Drug Administration, and American Heart Association; author's joint appointments at Johns Hopkins University and in listed industry positions have been reviewed and approved by Johns Hopkins University in accordance with its conflict-of-interest policies. **J.J.** No relevant relationships. **C.T.L.** No relevant relationships. **P.H.Y.** No relevant relationships. **C.M.H.** No relevant relationships.

References

1. Milam ME, Koo CW. The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. Clin Radiol 2023;78(2):115–122.
2. Kelly BS, Judge C, Bollard SM, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 2022;32(11):7998–8007. [Published correction appears in Eur Radiol 2022;32(11):8054.]
3. Mello-Thoms C, Mello CAB. Clinical applications of artificial intelligence in radiology. Br J Radiol 2023;96(1150):20221031.
4. Strohm L, Hehakaya C, Ranschaert ER, Boon WPC, Moors EHM. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. Eur Radiol 2020;30(10):5525–5532.
5. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit Med 2021;4(1):31.
6. Henry KE, Kornfield R, Sridharan A, et al. Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit Med 2022;5(1):97.
7. Eng DK, Khandwala NB, Long J, et al. Artificial intelligence algorithm improves radiologist performance in skeletal age assessment: a prospective multicenter randomized controlled trial. Radiology 2021;301(3):692–699.
8. Han SS, Kim YJ, Moon IJ, et al. Evaluation of artificial intelligence–assisted diagnosis of skin neoplasms: a single-center, paralleled, unmasked, randomized controlled trial. J Invest Dermatol 2022;142(9):2353–2362.e2.
9. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized controlled trials of artificial intelligence in clinical practice: systematic review. J Med Internet Res 2022;24(8):e37188.
10. Adams R, Henry KE, Sridharan A, et al. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat Med 2022;28(7):1455–1460.
11. Dratsch T, Chen X, Rezazade Mehrizi M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 2023;307(4):e222176.
12. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: Proceedings of the 4th Machine Learning for Healthcare Conference. Proc Mach Learn Res 2019;106:359–380.
13. Gastounioti A, Kontos D. Is it time to get rid of black boxes and cultivate trust in AI? Radiol Artif Intell 2020;2(3):e200088.
14. Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiol Artif Intell 2020;2(3):e190043.
15. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. Lancet 2022;399(10325):620.
16. van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;79:102470.
17. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. Comput Biol Med 2022;140:105111.
18. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv 1702.08608 [preprint] https://arxiv.org/abs/1702.08608. Posted February 28, 2017. Updated March 2, 2017. Accessed October 24, 2024.
19. Jair Escalante H, Escalera S, Guyon I, et al. Explainable and interpretable models in computer vision and machine learning. Springer, 2018.
20. Kim E, Kim S, Seo M, Yoon S. XProtoNet: diagnosis in chest radiography with global and local explanations. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers, 2021; 15714–15723.
21. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit Med 2021;4(1):4.
22. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021;3(11):e745–e750.
23. Chen H, Gomez C, Huang CM, Unberath M. Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. NPJ Digit Med 2022;5(1):156.
24. Gaube S, Suresh H, Raue M, et al. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. Sci Rep 2023;13(1):1383.
25. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv 1406.5823 [preprint] https://arxiv.org/abs/1406.5823. Posted June 23, 2014. Accessed October 24, 2024.
26. R Core Team. R: a language and environment for statistical computing. Version 4.2.0. Vienna, Austria: R Foundation for Statistical Computing, 2018.
27. Ehrmann DE, Gallant SN, Nagaraj S, et al. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nat Med 2022;28(7):1331–1333.
28. Parikh JR, Wolfman D, Bender CE, Arleo E. Radiologist burnout according to surveyed radiology practice leaders. J Am Coll Radiol 2020;17(1):78–81.
29. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Madsen MT, Kramer DJ. Do long radiology workdays affect nodule detection in dynamic CT interpretation? J Am Coll Radiol 2012;9(3):191–198.
30. Finck T, Moosbauer J, Probst M, et al. Faster and better: how anomaly detection can accelerate and improve reporting of head computed tomography. Diagnostics (Basel) 2022;12(2):452.
31. Lam Shin Cheung J, Ali A, Abdalla M, Fine B. U"AI" testing: user interface and usability testing of a chest X-ray AI tool in a simulated real-world workflow. Can Assoc Radiol J 2023;74(2):314–325.
32. Sivaraman V, Bukowski LA, Kahn JM, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In: CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, 2023.
33. Ferrario A, Loi M, Viganò E. In AI we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. Philos Technol 2020;33(3):523–539.
34. Ferrario A, Loi M, Viganò E. Trust does not need to be human: it is possible to trust medical AI. J Med Ethics 2020;47(6):437–438.