Journal of the American Medical Informatics Association, 29(12), 2022, 2041–2049
https://doi.org/10.1093/jamia/ocac179
Advance Access Publication Date: 13 October 2022
Research and Applications

Research and Applications

How does the artificial intelligence-based image-assisted technique help physicians in diagnosis of pulmonary adenocarcinoma? A randomized controlled experiment of multicenter physicians in China

Jiaoyang Li[1], Lingxiao Zhou[2], Yi Zhan[3], Haifeng Xu[4], Cheng Zhang[5], Fei Shan[3], and Lei Liu[6]

[1]School of Business Administration, Faculty of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China, [2]Institute of Microscale Optoelectronics, Shenzhen University, Shenzhen 518060, China, [3]Department of Radiology, Shanghai Public Health Clinical Center, Fudan University, Shanghai 201508, China, [4]Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China, [5]School of Management, Fudan University, Shanghai 200433, China, and [6]Intelligent Medicine Institute, Fudan University, Shanghai 200030, China

Corresponding Author: Cheng Zhang, PhD, School of Management, Fudan University, No 670, Guoshun Road, Yangpu District, Shanghai 200433, China; zhangche@fudan.edu.cn

Received 8 July 2022; Revised 24 August 2022; Editorial Decision 17 September 2022; Accepted 24 September 2022

ABSTRACT

Objective: Although artificial intelligence (AI) has achieved high levels of accuracy in the diagnosis of various diseases, its impact on physicians' decision-making performance in clinical practice is uncertain. This study aims to assess the impact of AI on the diagnostic performance of physicians with differing levels of self-efficacy under working conditions involving different time pressures.

Materials and methods: A 2 (independent diagnosis vs AI-assisted diagnosis) $\times$ 2 (no time pressure vs 2-minute time limit) randomized controlled experiment of multicenter physicians was conducted. Participants diagnosed 10 pulmonary adenocarcinoma cases and their diagnostic accuracy, sensitivity, and specificity were evaluated. Data analysis was performed using multilevel logistic regression.

Results: One hundred and four radiologists from 102 hospitals completed the experiment. The results reveal (1) AI greatly increases physicians' diagnostic accuracy, either with or without time pressure; (2) when no time pressure, AI significantly improves physicians' diagnostic sensitivity but no significant change in specificity, while under time pressure, physicians' diagnostic sensitivity and specificity are both improved with the aid of AI; (3) when no time pressure, physicians with low self-efficacy benefit from AI assistance thus improving diagnostic accuracy but those with high self-efficacy do not, whereas physicians with low and high levels of self-efficacy both benefit from AI under time pressure.

Discussion: This study is one of the first to provide real-world evidence regarding the impact of AI on physicians' decision-making performance, taking into account 2 boundary factors: clinical time pressure and physicians' self-efficacy.

Conclusion: AI-assisted diagnosis should be prioritized for physicians working under time pressure or with low self-efficacy.

Key words: artificial intelligence, multicenter physicians, diagnostic performance, time pressure, self-efficacy

BACKGROUND AND SIGNIFICANCE

With the rapid development of artificial intelligence (AI) in the medical field over the last decade (eg,[1–4]), there has been an emerging trend for AI technology to assist physicians in their diagnoses.[5] A promising approach is human–AI collaboration, in which AI serves as an aid to augment physicians' decision-making capabilities.[6] Before applying human–AI collaboration to a wide range of clinical practice, it is imperative to assess the impact of AI on physicians' medical decision-making quality in the real world.

Most of the existing research on medical AI has focused on technical aspects, aiming to develop more accurate and transparent medical AI algorithms,[7] yet the implementation of AI in real-world clinical settings is still in its infancy and lacks valid evidence.[8] Moreover, mixed outcomes have been found in the sparse literature evaluating AI in clinical practice. For example, some researchers have found that physicians achieve better performance in disease diagnosis when augmented by AI algorithms.[9,10] In contrast, other studies did not find evidence that AI-enhanced physicians' decision-making performance.[11,12] This may be because some studies used small or trainee samples,[13] or because the role of AI in medical decision-making varies with clinical settings and decision-making subjects. However, to date, there is a dearth of field studies exploring the effects of human–AI collaboration with multicenter physician samples,[14] and little is known about which physician groups could benefit more from medical AI. A comprehensive and robust evaluation of these issues is key to advancing medical AI from theory to clinical practice.[13] Additionally, obtaining answers requires strong validation in real-world samples of physicians. This study aims to fill these gaps.

We explore the impact of AI on physician decision-making by conducting a randomized controlled experiment of multicenter physicians in a real-world clinical scenario of pulmonary adenocarcinoma diagnosis through reading CT images. We chose this scenario for the following reasons: on the one hand, it is a challenging task for physicians to distinguish between different subtypes of pulmonary mininodules through reading CT images,[15] and clinical management strategies vary for different subtypes of adenocarcinoma. Adenocarcinomas in situ and minimally invasive adenocarcinomas (minimally/noninvasive nodules) can be clinically followed up or undergo limited resection to reduce overtreatment, while invasive adenocarcinomas (invasive nodules) require timely lobectomy and mediastinal lymph node dissection.[16,17] Therefore, the classification of pulmonary nodules through CT images is important for the selection of appropriate clinical decision-making strategies.[17] And on the other hand, AI techniques for classifying pulmonary adenocarcinoma nodules through CT images can achieve a better performance beyond that of physicians.[18] We not only focus on the role of AI in the overall diagnostic accuracy of physicians but also scrutinize 3 subresearch questions: (1) the role of AI in a time-pressured environment—a crucial working condition in clinical practice;[19] (2) the role of AI for physicians' diagnostic sensitivity and specificity; and (3) the role of AI in the diagnostic accuracy of physicians with different levels of self-efficacy. Answering these questions is key to use AI right before the large-scale implementation of AI.[20]

MATERIALS AND METHODS

System preparation

Prior to the experiment, an AI-assisted pulmonary adenocarcinoma diagnosis system was developed in cooperation with Zhongshan Hospital and Shanghai Public Clinical Health Center, 2 of China's top hospitals. This diagnosis system differs from previous systems used by radiologists in that it displays AI-based diagnoses to the radiologists as an aid (Supplementary Appendix Figure A2).

The AI-assisted diagnosis system has an interactive interface running in a web browser for radiologists to read the pulmonary CT images with nodular lesions and make diagnoses. For each case, the patient's pulmonary CT images (30 two-dimensional nodule-centered CT axial slices, containing 15 slices each above and below centered on the nodule in question; a series of consecutive CT images up to 1 mm thick, size = 512×512 pixels), the number and location of nodules, and demographic information including age, gender, smoking history, and tumor history were displayed, on the basis of which the radiologists perform a binary classification of diagnosis as to whether the patient's pulmonary nodules are invasive or minimally/noninvasive (Supplementary Appendix Figure A1). Minimally/noninvasive nodules have good biological behavior with no long-term changes or slow growth, allowing for selection of the optimal surgical time point through clinical follow-up, and even if surgery is performed, it is a limited sublobar resection that can preserve more lung function, reduce postoperative complications, and shorten recovery time.[16,17] In contrast, invasive nodules require lobectomy and mediastinal lymph node dissection and postoperative adjuvant therapy, which may improve survival rates.[16,17] Therefore, effective evaluation of patients with pulmonary nodules can expedite the treatment of invasive nodules and reduce the overtreatment of patients with minimally/noninvasive nodules.[21] The system provides various image-reading functions such as positioning, zooming in and out, measuring, flipping, and multiplanar reconstruction including sagittal and coronal.

The AI-based diagnostic suggestions were derived from the algorithm previously developed by our team members to distinguish invasive versus minimally/noninvasive from subcentimeter pulmonary adenocarcinomas.[18] The algorithm was generated by a generative adversarial network (GAN)-based image augmentation method with progressive growing and pixel-wise normalization, followed by a convolutional neural network (CNN) with 4 convolution layers, 4 max-pooling layers, and 1 fully connected layer.[18] The GAN method can improve the classification performance of CNN while alleviating the problem of insufficient medical image datasets.[18] Through training with 206 pulmonary nodules with postoperative pathological labels, the algorithm achieved an accuracy of 80.03% on invasive versus minimally/noninvasive prediction, which was comparable to state-of-the-art methods.[18] Therefore, the results were used as AI suggestions to radiologists in our AI-assisted diagnosis system.

Participants recruitment

We recruited radiologist participants across China through a professional medical information service platform (https://www.medlive.cn/). Radiologists registered on this platform have been certified in terms of their workplace and department. An experiment invitation with a participating link was sent to certified radiologists via email and SMS (refer to Supplementary Appendix Table A1 for details). The radiologists who accepted the invitation were required to complete a screening questionnaire that included their department and CT diagnosis experience. Those who did not work in the radiology/imaging department or who had no experience in CT diagnosis were excluded (Supplementary Appendix Table A2) from the follow-up experiment to ensure that all participants had the basic ability to

diagnose pulmonary adenocarcinoma through CT images. The recruitment lasted from July to September 2020.

Experiment design

We first invited 7 radiologists from Zhongshan Hospital and Shanghai Public Clinical Health Center to conduct a pilot study. They all specialized in cardiothoracic conditions, with CT diagnosis experience ranging from less than 10 years to more than 20 years. The average time they spent diagnosing each case was about 2.5 minutes, and based on their suggestions, we believe that a 2-minute time limit is an appropriate setting for physicians to feel pressure during the diagnostic process. In actual clinical practice, time pressure is also a common situation for radiologists,[22] who need to interpret a CT slice in 3–4 seconds to meet workload requirements (in our experimental scenario, each case contains 30 CT axial slices, which need to be interpreted in 1.5–2 min).[23] Therefore, the 2-minute time pressure is a realistic setting. We also refined the details of the experiment based on their feedback to ensure that the experiment was close to a real clinical practice scenario.

Then, we employed a 2 (independent diagnosis vs AI-assisted diagnosis) × 2 (no time pressure vs 2-minute time limit) between-subjects design. The qualified participants were randomly assigned to 1 of the 4 conditions using a computerized random number generator. Figure 1 shows the process for the recruitment and experiment.

Experiment procedure

In the beginning, all participants were asked to report their age, gender, and self-efficacy level and then learned how to use the diagnosis system by reading the user guide (Supplementary Appendix Table A3). Self-efficacy refers to a person's belief in his or her capability to successfully perform a particular task.[24] Such beliefs may influence the extent to which they adopt and benefit from AI suggestions. In our experiment, we measured the radiologists' self-efficacy level by asking about the accuracy they thought they could achieve in diagnosing pulmonary adenocarcinoma (invasive vs minimally/noninvasive).[25] Physicians above the average were considered to have high self-efficacy, while those below the average were considered to have low self-efficacy. They were allowed to refer to this guide at any time during the diagnostic process by clicking on the "User Guide" button. Participants in the AI-assisted groups additionally read a brief introduction to the AI technique and its applications (Supplementary Appendix Table A4) and were informed that it was 80% accurate in classifying invasive versus minimally/noninvasive cases, which was consistent with the actual performance of the AI suggestions.

Next, all participants were asked to diagnose pulmonary adenocarcinoma cases using the diagnosis system (Supplementary Appendix Figure A1). For each case, participants had to complete a 2-step diagnostic process. First, they were required to make an initial diagnosis (invasive vs minimally/noninvasive) on their own. Upon submission, the radiologists in the AI-assisted groups were shown AI-based suggestions, while those in the independent diagnosis groups were not. Second, all participants were prompted to confirm the diagnosis. At this point, participants could choose to stick with or change their initial decision, and then submit the final diagnosis by clicking the "Confirm" button (Supplementary Appendix Figure A2). Radiologists in the no time pressure groups were not imposed a time limit for diagnosing each case, while those in the time pressure groups had a 2-minute time limit for diagnosis of each case with a countdown timer reminder. This 2-submission decision process helped to avoid automation bias; that is, presenting automated cues (AI suggestions) to aid users before their decision-making process may lead to clinical over-reliance on automation.[26,27]

Each radiologist was required to diagnose a total of 13 cases throughout the experiment, with the first 3 being a warm-up session to familiarize themselves with the system and the last 10 being a formal experimental session. A case database was constructed by 2 senior radiologists with more than 20 years of pulmonary adenocarcinoma CT diagnosis experience through selecting from 206 real clinical cases. The criteria for case selection were that both radiologists agreed that the cases had average or above average diagnostic difficulty based on their experience. Ten cases were then randomly selected from the case database for the formal experiment using a computerized random number generator. Of the 10 cases, 5 were invasive and the other 5 were minimally/noninvasive based on postoperative pathological evaluations, a golden standard in clinical practice. The AI could only correctly diagnose 4 of invasive cases and 4 of minimally/noninvasive cases with an accuracy rate of 80%, which was consistent with the actual performance of our AI algorithm.

The experiment then ended with a payment based on each participant's performance. The entire experiment was estimated to last 30 to 40 minutes. Radiologists who completed the whole process received a base payment of 200 RMB (equivalent to approximately 30 USD), and a bonus of 20 RMB (approximately 3 USD) for each correct final diagnosis, which was intended to motivate them to be more conscientious and engaged and to put more effort into the diagnostic process. The experiment was approved by Shanghai public health clinical center ethics committee (2020-S139-01).

RESULTS

Participant information

One hundred and four multicenter physicians were enrolled from 102 hospitals in China. Details of the hospital name, region, and level are shown in Supplementary Appendix Table A5. In total, there were 88 male (84.62%) and 16 female participants (15.38%), with an average age of 43.76 years (SD = 7.26 years). 23 participants had less than 10 years of experience in CT diagnosis (22.12%), 47 had CT experience between 10–20 years (45.19%), and 34 had CT experience of more than 20 years (32.69%). The Chi-square test showed no statistically significant differences in gender distribution between the 4 groups ($P = .385$). The Kruskal–Wallis test suggested no significant differences in terms of age ($P = .812$) and experience ($P = .958$) across the 4 groups.

Manipulation check

We compared the mean time spent by radiologists on diagnosing 10 cases between the groups manipulated by different time pressure conditions. It was found that the average time spent by radiologists in the 2 groups with 2-minute time limit was significantly shorter than in the 2 groups under no time pressure ($P = .021$), suggesting that the manipulation of time pressure was effective.

Outcome

We performed data analysis using multilevel logistic regression models that incorporated groupings by case. This approach allows us to assess the impact of AI assistance on the probability of radiologists making correct diagnosis while controlling for radiologists' gender and years of CT diagnosis experience. The tool used was the *melogit* command in the STATA software. (https://www.stata-press.com/

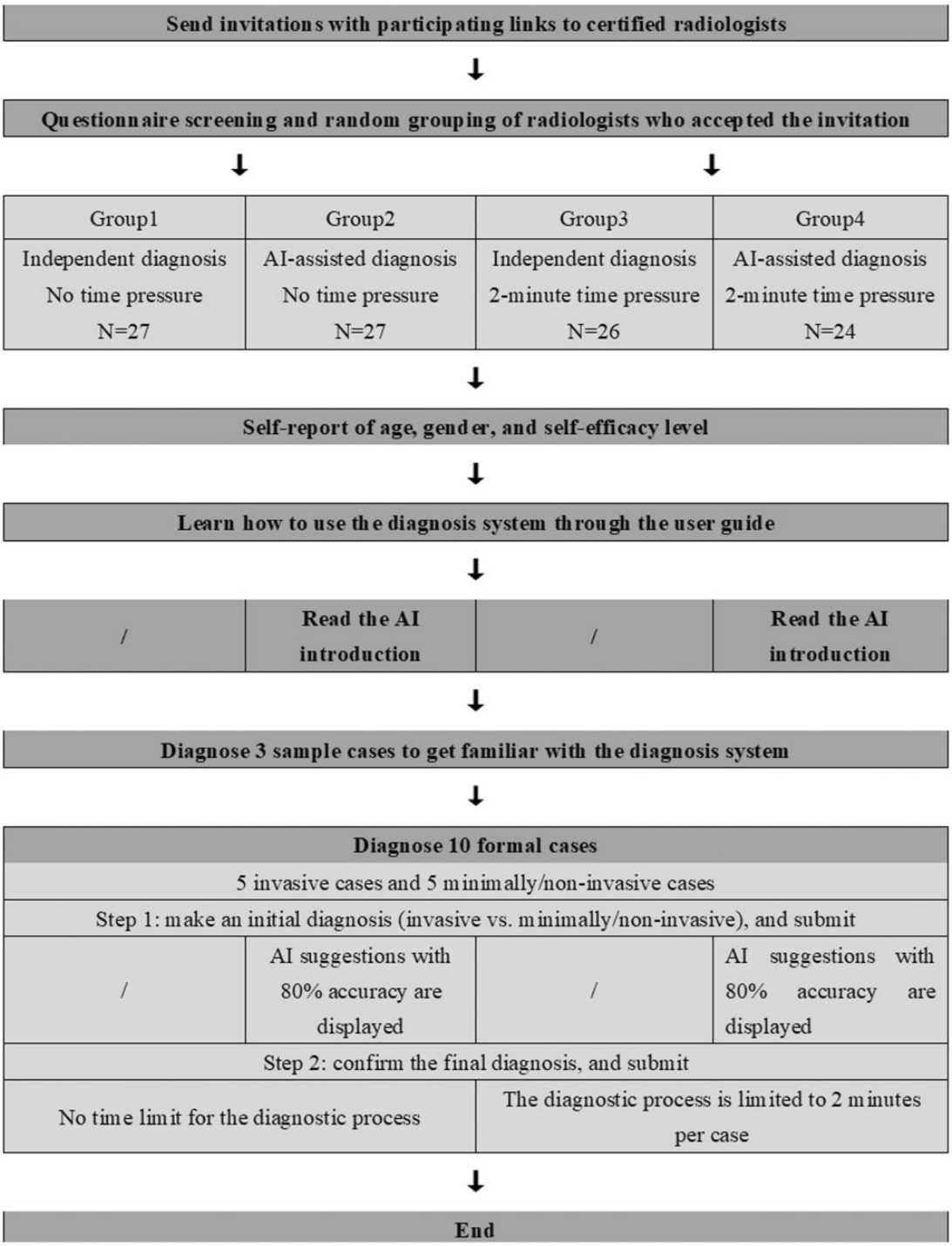

Figure 1. The process of participant recruitment and experiment.

manuals/multilevel-mixed-effects-reference-manual/, accessed 12 August 2022) The results are presented in Table 1. We also used the *prtest* command in the STATA software to calculate the mean accuracy, sensitivity, specificity, and their confidence intervals (CI) below.[28]

First, we examined whether AI is helpful to physicians in different work environments. With no time limit, the average accuracy of radiologists' independent diagnosis was 60.37% (95% CI: 54.54%–66.20%), and that of radiologists assisted by AI was 70.74% (95% CI: 65.31%–76.17%). The odds ratio of 1.58 ($P = .021$) indicates that radiologists with AI support were more likely to make correct diagnosis than those without, which means AI-assisted diagnosis could improve radiologists' diagnostic accuracy.

Under 2-minute time pressure, the average accuracy of radiologists' independent diagnosis (55.38%, 95% CI: 49.34%–61.43%) was slightly lower than that without time pressure (a decrease of 4.99%). With the aid of AI, the diagnostic accuracy of radiologists under time pressure was improved to 70.42% (95% CI: 64.64%–

Table 1. Multilevel logistic regression analysis of radiologists' diagnoses by specific intervention type: with versus without AI assistance

	No time pressure		2-minute time pressure	
	OR (SE)	*P*-value	OR (SE)	*P*-value
Probability of correct diagnoses for all cases (Accuracy)				
Intercept	2.88 (1.21)	.012*	0.81 (0.31)	.588
AI assistance	1.58 (0.31)	.021*	2.13 (0.44)	.000***
Probability of correct diagnoses for invasive cases (Sensitivity)				
Intercept	4.96 (3.43)	.021*	0.80 (0.50)	.725
AI assistance	2.45 (0.72)	.002**	2.20 (0.69)	.012*
Probability of correct diagnoses for minimally/noninvasive cases (Specificity)				
Intercept	2.06 (1.09)	.171	0.80 (0.34)	.602
AI assistance	1.09 (0.30)	.760	2.08 (0.57)	.007**
Probability of correct diagnoses for all cases by radiologists with low self-efficacy (Accuracy of radiologists with low self-efficacy)				
Intercept	2.48 (1.13)	.046*	0.79 (0.35)	.590
AI assistance	1.55 (0.34)	.045*	2.02 (0.48)	.004**
Probability of correct diagnoses for all cases by radiologists with high self-efficacy (Accuracy of radiologists with high self-efficacy)				
Intercept	5.43 (6.73)	.173	0.95 (0.55)	.923
AI assistance	1.35 (0.81)	.615	2.80 (1.51)	.055+

Note: Controlling for gender and years of experience in CT diagnosis.
***$P < .001$; **$P < .01$; *$P < .05$; +$P < .1$. OR: odds ratio; SE: standard error.

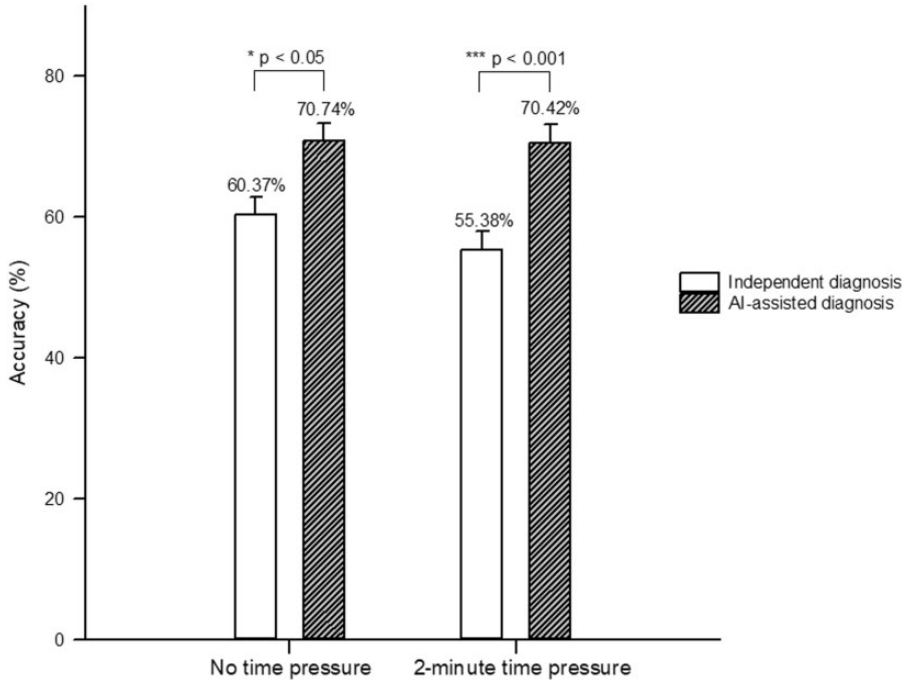

Figure 2. The average accuracy of radiologists' independent and AI-assisted diagnoses.

76.19%). The odds ratio of 2.13 ($P < .001$) suggests that physicians could benefit from AI-assisted diagnosis in time-pressured environment as well. Figure 2 presents the average accuracy of radiologists' independent and AI-assisted diagnoses in a work environment with and without time pressure.

Second, after establishing that AI can indeed help physicians improve diagnostic accuracy, we scrutinized whether AI contributes to physicians' diagnostic sensitivity or specificity. Sensitivity refers to the true positive (true invasive) rate, while specificity reflects the true negative (true minimally/noninvasive) rate. Under no time pressure, radiologists had an average sensitivity of 54.07% (95% CI: 45.67%–62.48%) for independent diagnosis and 71.85% (95% CI: 64.27%–79.44%) for AI-assisted diagnosis. The odds ratio of 2.45 ($P = .002$) shows that radiologists were better able to correctly diagnose invasive cases with the aid of AI than without, that is, AI-assisted diagnosis could increase radiologists' diagnostic sensitivity. However, AI assistance did not significantly impact the radiologists' diagnostic specificity level (independent diagnosis = 66.67%, 95% CI: 58.71%–74.62%; AI-assisted diagnosis = 69.63%, 95% CI: 61.87%–77.39%; odds ratio = 1.09, $P = .760$).

Compared to the setting without time pressure, there was no significant change in the average sensitivity of radiologists' independ-

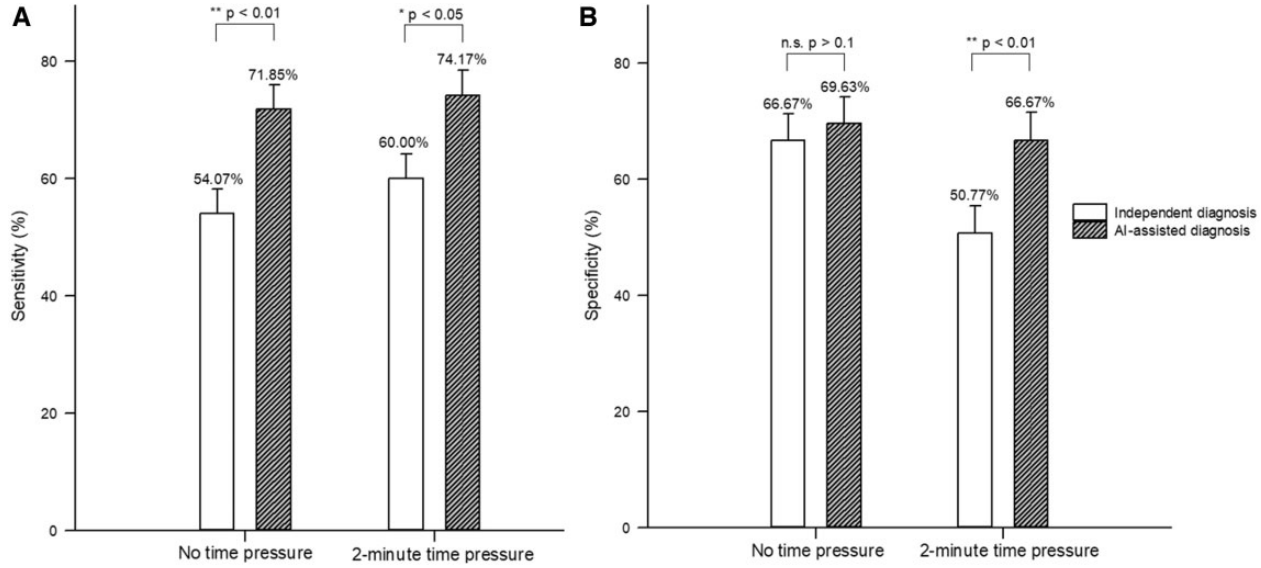

Figure 3. The average sensitivity (A) and specificity (B) of radiologists' independent and AI-assisted diagnoses.

ent diagnosis (60.00%, 95% CI: 51.58%–68.42%) under 2-minute time pressure (an increase of 5.93%), while the average specificity of their independent diagnosis (50.77%, 95% CI: 42.18%–59.36%) declined substantially (a decrease of 15.90%). With the aid of AI, radiologists were able to achieve a sensitivity of 74.17% (95% CI: 66.34%–82.00%) and a specificity of 66.67% (95% CI: 58.23%–75.10%). The multilevel logistic regression result implies that physicians benefited from AI to simultaneously increase sensitivity (odds ratio = 2.20, $P = .012$) and specificity (odds ratio = 2.08, $P = .007$) in time-critical settings. Figure 3A and B displays the average sensitivity and specificity of the radiologists' independent and AI-assisted diagnoses with and without time pressure.

Third, we further explored whether the benefits of AI-assisted diagnosis are heterogeneous for physicians with different levels of self-efficacy. In the absence of time pressure, AI-assisted diagnosis significantly enhanced the average diagnostic accuracy of radiologists with low self-efficacy (independent diagnosis = 58.26%, 95% CI: 51.89%–64.63%; AI-assisted diagnosis = 68.50%, 95% CI: 62.06%–74.94%; odds ratio = 1.55, $P = .045$). However, radiologists with high self-efficacy did not benefit significantly from AI (independent diagnosis = 72.50%, 95% CI: 58.66%–86.34%; AI-assisted diagnosis = 77.14%, 95% CI: 67.31%–86.98%; odds ratio = 1.35, $P = .615$).

Under 2-minute time pressure, the independent diagnostic accuracy of radiologists with low self-efficacy (56.88%, 95% CI: 49.20%–64.55%) did not change significantly compared with that under no time pressure (a decrease of 1.38%), whereas that of radiologists with high self-efficacy (53.00%, 95% CI: 43.22%–62.78%) dropped a lot (a decrease of 19.50%). This implies that time pressure severely impairs the decision-making quality of physicians with high self-efficacy. With the assistance of AI, radiologists with low self-efficacy could achieve a substantial increase in diagnostic accuracy (70.48%, 95% CI: 64.31%–76.65%; odds ratio = 2.02, $P = .004$). More fortunately, radiologists with high self-efficacy could also leverage AI to make up for the loss of diagnostic accuracy due to time pressure (70.00%, 95% CI: 53.60%–86.40%; odds ratio = 2.80, $P = .055$), although it was marginally significant. Figure 4A and B presents the average accuracy of independent and AI-assisted diagnoses for radiologists with low (high) self-efficacy in a work environment with and without time pressure context.

DISCUSSION

Although medical AI technology has rivaled or surpassed medical experts in many areas, it is unclear how it affects physicians' diagnostic performance in clinical practice, and limited studies have yielded mixed results.[9–12] Through a randomized controlled experiment of multicenter physicians, this study investigated the impact of AI on physicians' diagnostic performance in the context of pulmonary adenocarcinoma diagnosis, taking into account 2 boundary factors: clinical work environment and physicians' self-efficacy.

First, we considered time pressure as an important clinical work environment. Due to the personnel shortage in the medical field, time pressure is a common working condition in clinical decision-making.[19,29] Previous studies suggest that excessive time pressure can lead to information overload, undermining the quality of decisions,[30,31] and increasing diagnostic errors.[22] Therefore, the impact of AI assistance on physicians' diagnostic performance under high time pressure is worthy of attention, but this critical work environment has been overlooked in relevant studies. Our results show that AI assistance can greatly improve physicians' diagnostic accuracy to around 70%, with or without time pressure.

Second, in addition to accuracy, we also assessed the sensitivity and specificity metrics. In many disease diagnostic situations, the cost of undetected disease outweighs the cost of overprescribing treatment or further testing; therefore false positive results are prevalent.[32,33] Techniques are needed to reduce false positive results (ie, enhance specificity) while maintaining high sensitivity.[34] Some studies have shown that AI can improve both the sensitivity and specificity of physicians' diagnosis,[35,36] while others have found that AI only improves sensitivity but not specificity.[37] We find that under no time pressure, AI assistance significantly increases physicians' diagnostic sensitivity but the specificity does not change significantly. In contrast, under the 2-minute time limit, both the physicians' diagnostic sensitivity and specificity are significantly

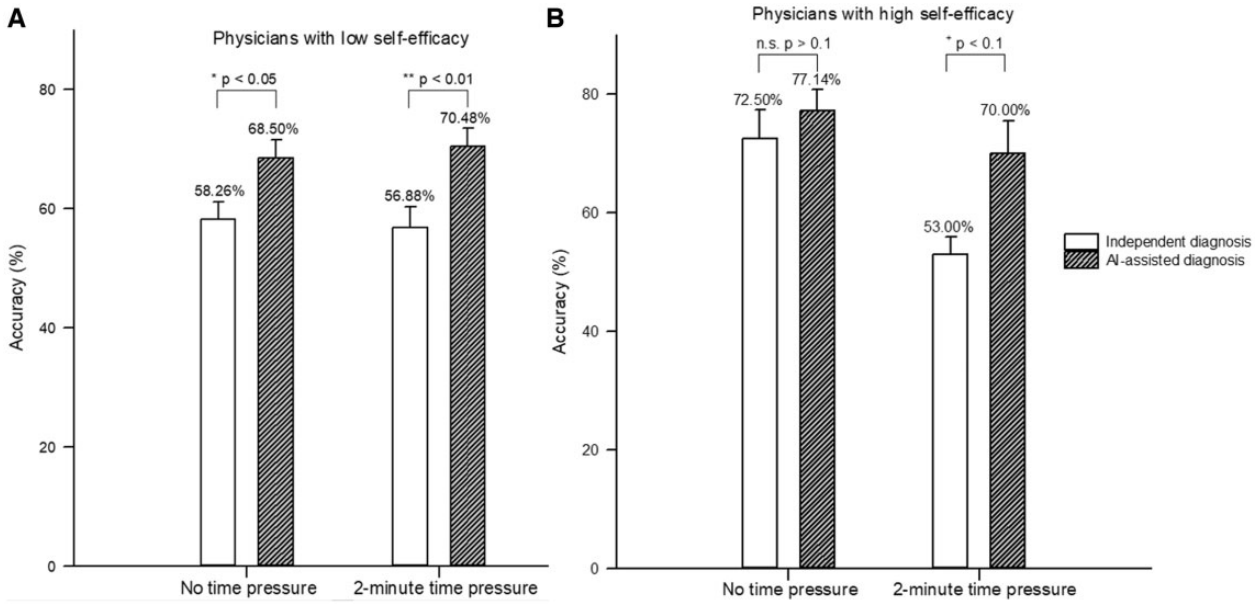

Figure 4. The average accuracy of independent and AI-assisted diagnoses for radiologists with low (A) or high (B) self-efficacy.

improved with the help of AI. This further demonstrates the effectiveness of the collaboration between physicians and AI, especially in the case of high time pressure, which can effectively improve specificity, that is, reduce the problem of false positives, while maintaining the high sensitivity of disease diagnosis.

Third, we further considered heterogeneity among physicians and evaluated which groups of physicians would benefit more from AI. Previous studies have shown that AI-assisted diagnosis should benefit those physicians who are most likely to make errors, such as junior physicians with limited experience.[5,6] However, in addition to objective experience, physicians' subjective perceptions often influence the extent to which they adopt and benefit from AI-based suggestions, which has received little attention before. We find that in the absence of time pressure, AI can significantly improve the diagnostic accuracy of physicians with low self-efficacy, but physicians with high self-efficacy do not benefit significantly because their own independent diagnostic accuracy is already high. When faced with time pressure, physicians of all levels of self-efficacy have a decreased accuracy rate for independent diagnosis, which fortunately can be greatly improved with the help of AI.

Our findings provide implications for clinical practice. First, physicians should actively adopt AI-assisted diagnostic systems in their clinical work and collaborate with AI to reach new workflows. While the previous workflow was that physicians made their own diagnoses to reach a conclusion, the current workflow is that when the diagnosis result cannot be determined, the physicians should refer to the AI suggestion before drawing a conclusion. This helps to improve the diagnostic performance of physicians, so that invasive patients can receive timely surgical resection to improve survival rate, and minimally/noninvasive patients can avoid excessive treatment (unnecessary puncture and surgery, etc.) to reduce physical and mental suffering. Second, AI-assisted diagnosis systems should be preferentially used by physicians in a working environment where medical resources are scarce and time pressure is high, so as to make up for the loss of diagnostic accuracy caused by time pressure and avoid the damage to patients' physical and mental health caused by misdiagnosis. Third, AI-assisted diagnostic systems should also be prioritized for those physicians with low self-efficacy, which can greatly improve their diagnostic accuracy and enable patients to receive the most timely and appropriate clinical management solutions to increase patient well-being.

There are still some limitations in this study. First, the study focused on a single clinical scenario——pulmonary adenocarcinoma, may limit the generalizability of our results to other contexts. Therefore, future studies should be conducted to validate the findings in various clinical decision-making contexts. Second, our research subjects were limited to Chinese radiologists, however, different cultural backgrounds may lead to different findings. Thus, future studies should expand the sample size and extend to other cultural contexts. Third, although the setting of the 2-minute time pressure and the AI suggestions with an accuracy of 80% are well-founded, slight differences in time constraints and AI accuracy may lead to quite different results. So, future studies should further explore the robustness of the findings under different time pressure and AI accuracy settings. Fourth, although we did our best to simulate real-world clinical decision-making in our experiments, there are still some gaps with reality. For example, the 10 cases we selected may not represent the full picture of cases radiologists encounter in their clinical work, which could be addressed by collecting more representative cases in future research. Additionally, the simulated environment may make the physician feel not truly responsible for the diagnosis and not behave as in a real clinical setting, so we tried to incentivize their engagement and effort through a pay-per-performance approach. Fifth, the study only explored 1 design for AI-assisted diagnosis while there are various designs. For instance, more information about the algorithm and outcome could be provided to explain AI suggestions, and different presentation formats of AI suggestions, such as text and images, may also affect physicians' decisions. While the study sheds some initial light on the human–AI collaborative decision in the clinical practice, more in-depth investigations on the human–AI design are encouraged to conduct in the future.

CONCLUSION

This study is one of the first to provide real-world evidence regarding the impact of AI on physicians' decision-making performance. It examines the impact of AI on the diagnostic performance of physicians with different levels of self-efficacy under working conditions involving different time pressures. The results show that the AI-based image-assist technique can significantly enhance the physicians' diagnostic performance either with or without time pressure and is most effective in helping low-efficacy physicians. The findings offer practical implications for both frontline physicians and patients.

FUNDING

National Natural Science Foundation of China (72101211, 81501294, 91846302, and 71871065); and the Clinical Research Plan of SHDC (SHDC2020CR3080B).

AUTHOR CONTRIBUTIONS

FS, CZ, LL, LXZ, and JYL designed the study; LXZ and LL implemented the system; YZ provided CT imaging; CZ, JYL, and HFX designed the experiment; JYL, LXZ, and FS collected data; JYL and CZ analyzed the data; JYL, CZ, FS, and LL interpreted the results; JYL, HFX, and CZ wrote and revised the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. Muse ED, Topol EJ. Guiding ultrasound image capture with artificial intelligence. *Lancet* 2020; 396 (10253): 749.
2. Attia ZI, Noseworthy PA, Lopez-Jimenez F, *et al.* An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *Lancet* 2019; 394 (10201): 861–7.
3. Milea D, Najjar RP, Zhubo J, *et al.*; BONSAI Group. Artificial intelligence to detect papilledema from ocular fundus photographs. *N Engl J Med* 2020; 382 (18): 1687–95.
4. Shen J, Zhang CJP, Jiang B, *et al.* Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Med Inform* 2019; 7 (3): e10010.
5. Jussupow E, Spohrer K, Heinzl A, *et al.* Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inform Syst Res* 2021; 32 (3): 713–35.
6. Tschandl P, Rinner C, Apalla Z, *et al.* Human–computer collaboration for skin cancer recognition. *Nat Med* 2020; 26 (8): 1229–34.
7. Rai A. Explainable AI: from black box to glass box. *J Acad Mark Sci* 2020; 48 (1): 137–41.
8. Nagendran M, Chen Y, Lovejoy CA, *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; 368: m689.
9. Meyer J, Khademi A, Têtu B, *et al.* Impact of artificial intelligence on pathologists' decisions: an experiment. *J Am Med Inform Assoc* 2022; 29 (10): 1688–95.
10. Rajpurkar P, O'Connell C, Schechter A, *et al.* CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest X-rays in patients with HIV. *NPJ Digit Med* 2020; 3 (1): 1–8.
11. Kiani A, Uyumazturk B, Rajpurkar P, *et al.* Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med* 2020; 3 (1): 23–8.
12. Jacobs M, Pradier MF, McCoy TH, *et al.* How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry* 2021; 11 (1): 1–9.
13. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res* 2021; 23 (4): e25759.
14. Angus DC. Randomized clinical trials of artificial intelligence. *JAMA* 2020; 323 (11): 1043–5.
15. Sun Z, Ng KH, Vijayananthan A. Is utilisation of computed tomography justified in clinical practice? Part I: Application in the emergency department. *Singapore Med J* 2010; 51 (3): 200–6.
16. Jiang Y, Che S, Ma S, *et al.* Radiomic signature based on CT imaging to distinguish invasive adenocarcinoma from minimally invasive adenocarcinoma in pure ground-glass nodules with pleural contact. *Cancer Imaging* 2021; 21 (1): 1–14.
17. Wang X, Li Q, Cai J, *et al.* Predicting the invasiveness of lung adenocarcinomas appearing as ground-glass nodule on CT scan using multi-task learning and deep radiomics. *Transl Lung Cancer Res* 2020; 9 (4): 1397–406.
18. Wang Y, Zhou L, Wang M, *et al.* Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification. *Quant Imaging Med Surg* 2020; 10 (6): 1249–64.
19. Linzer M, Konrad TR, Douglas J, *et al.*; the Society of General Internal Medicine (SGIM) Career Satisfaction Study Group (CSSG). Managed care, time pressure, and physician job satisfaction: results from the physician worklife study. *J Gen Internal Med* 2000; 15 (7): 441–50.
20. Taddeo M, Floridi L. How AI can be a force for good. *Science* 2018; 361 (6404): 751–2.
21. Mazzone PJ, Lam L. Evaluating the patient with a pulmonary nodule: a review. *JAMA* 2022; 327 (3): 264–73.
22. Vosshenrich J, Brantner P, Cyriac J, *et al.* Quantifying radiology resident fatigue: analysis of preliminary reports. *Radiology* 2021; 298 (3): 632–9.
23. McDonald RJ, Schwartz KM, Eckel LJ, *et al.* The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad Radiol* 2015; 22 (9): 1191–8.
24. Bandura A. Self-efficacy. In: Ramachaudran VS, ed. *Encyclopedia of Human Behavior*. Vol. 4. New York: Academic Press; 1994: 71–81.
25. Maurer TJ, Pierce HR. A comparison of Likert scale and traditional measures of self-efficacy. *J Appl Psychol* 1998; 83 (2): 324–9.
26. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; 24 (2): 423–31.
27. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012; 19 (1): 121–7.
28. Park HM. Comparing group means: T-tests and one-way ANOVA using STATA, SAS, R, and SPSS. The University Information Technology Serviced (UTTS) Center for Statistical and Mathematical Computing, Indiana University; 2009. http://www.indiana.edu/~statmath/stat/all/ttest.
29. Stepanikova I. Racial-ethnic biases, time pressure, and medical decisions. *J Health Soc Behav* 2012; 53 (3): 329–43.
30. Paul S, Nazareth DL. Input information complexity, perceived time pressure, and information processing in GSS-based work groups: an experimental investigation using a decision schema to alleviate information overload conditions. *Decis Support Syst* 2010; 49 (1): 31–40.
31. Hwang MI. Decision making under time pressure: a model for information systems research. *Inform Manag* 1994; 27 (4): 197–203.

32. Luce MF, Kahn BE. Avoidance or vigilance? The psychology of false-positive test results. *J Consum Res* 1999; 26 (3): 242–59.
33. Liang W, Zhao Y, Huang W, *et al.* Non-invasive diagnosis of early-stage lung cancer using high-throughput targeted DNA methylation sequencing of circulating tumor DNA (ctDNA). *Theranostics* 2019; 9 (7): 2056–70.
34. Elmore JG, Barton MB, Moceri VM, *et al.* Ten-year risk of false positive screening mammograms and clinical breast examinations. *N Engl J Med* 1998; 338 (16): 1089–96.
35. Bai HX, Wang R, Xiong Z, *et al.* AI augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other etiology on chest CT. *Radiology* 2020; 296 (3): E156–65.
36. Conant EF, Toledano AY, Periaswamy S, *et al.* Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell* 2019; 1 (4): e180096.
37. Park A, Chute C, Rajpurkar P, *et al.* Deep learning-assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA Netw Open* 2019; 2 (6): e195600.