Beyond Augmentation: Immediate Learning from AI in Expert Decision Making

Guanmei Ruan[1] Fengyan Cai[1, 2] Yong Lu[3] Na Geng[2,4]* Susan Feng Lu[5]*

[1] Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China

[2] Data-Driven Management Decision Making Lab, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China

[3] Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, 200025, China

[4] Sino-US Global Logistics Institute, Shanghai Jiao Tong University, Shanghai 200030, China

[5] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada

*Correspondence: gengna@sjtu.edu.cn; susanfeng.lu@utoronto.ca

Abstract. Artificial intelligence (AI) is increasingly used to augment expert decision making, yet little is known about whether interacting with AI improves human capability beyond the period of use. We study this question using a multi-stage experiment involving radiologists diagnosing lung nodules from chest CT images. Physicians first diagnose cases without AI assistance, are then randomly assigned to diagnose cases with or without AI support and finally diagnose additional cases after AI assistance is removed. The results show that AI assistance improves contemporaneous performance, increasing diagnostic accuracy and confidence while reducing decision time. More importantly, physicians previously exposed to AI perform better even after AI assistance is removed, including on new cases, suggesting learning spillovers from AI exposure. We further find that these effects depend not only on physicians' capability but also on their self-evaluation of expertise: physicians with lower self-evaluation benefit significantly more from AI assistance and exhibit larger post-AI learning improvements. These findings suggest that AI decision support exhibits dynamic complementarity with human expertise. Beyond augmenting human performance, AI may function as a learning technology that enhances human decision capability over time.

Keywords: AI-assisted diagnosis; Physician heterogeneity; Learning Spillovers; Self-evaluation; Dynamic Complementarity

1. Introduction

Artificial Intelligence (AI) decision support is commonly evaluated by the quality and speed of decisions made while the tool is present. For a professional service organization, another operational consequence may matter: what happens to the knowledge worker's next decision when the tool is absent? When AI highlights task-relevant information, workers may adjust how they interpret future cases. Conversely, a tool that supplies an answer may leave unaided judgment unchanged or encourage reliance. These possibilities make the timing of evaluation central to the value of AI adoption (Noy and Zhang 2023, Brynjolfsson et al. 2025, Chassagnon et al. 2023).

In healthcare practice, AI systems often provide physicians with feature-level decomposition that reveals underlying patterns that cannot be easily captured by human eyes in complex decision problems, rather than merely generating predictions. Through repeated exposure to algorithmic recommendations, physicians may learn to recognize important features, refine diagnostic heuristics, and improve their own decision strategies. If this occurs, AI may function not only as a productivity-enhancing tool but also as a capability-building technology that upgrades human expertise. Understanding whether such learning spillovers exist is crucial for organizations adopting AI, because it informs whether AI serves as *a capability-building technology* or, alternatively, contributes to *the erosion of human expertise*. Despite the importance of this question, experimental evidence using professional knowledge workers such as medical providers remains limited in the literature.

We study immediate learning as a change in AI-unaided performance on subsequent cases. This definition is deliberately narrow. A gain on a new case after the tool is removed is consistent with transfer of a recently acquired cue or strategy; it does not establish durable skill acquisition, retention over days or months, or improved patient outcomes. The distinction matters in healthcare operations, where access to decision support may vary across shifts, facilities, and systems.

We address the research question using a large-scale, multi-stage experiment involving practicing radiologists performing lung nodule diagnostic tasks. The experimental design consists of three stages. In the first stage, physicians diagnose cases without AI assistance. In the second stage, physicians are randomly assigned to perform diagnoses on a different set of cases, either with or without AI support. In the final stage, AI assistance is removed for all physicians, allowing us to test whether prior exposure to AI affects subsequent unaided decision performance. The design strategically distinguishes learning

from task familiarity by including both repeated and new cases in the final stage. We evaluate performance along two dimensions central to operational decision making: quality and speed, captured by diagnostic accuracy and decision time, respectively.

Our main analytic sample consists of 321 physicians, each completing 20 case presentations. Relative to the comparison cohort, physicians in the AI-exposure cohort improved their graded diagnostic accuracy by 1.399 points, or 20.5%, while reducing decision time by 22% during AI access. Importantly, these gains persisted immediately after AI assistance was removed: the AI-exposure cohort exhibited a 1.479-point, or 21.6%, greater improvement in diagnostic accuracy and completed cases 20% faster than the comparison cohort in Stage 3. The post-exposure accuracy gain was 1.646 points for new cases and 1.311 points for repeated cases, and the two estimates were not statistically different from one another. Taken together, the results reveal a quality–speed double win: AI improves performance during use and is followed by higher accuracy and faster decisions after removal. Importantly, these gains extend to new cases, indicating an immediate transfer of newly acquired diagnostic knowledge beyond the cases encountered with AI. AI exposure also increases diagnostic confidence, suggesting an indirect channel through which AI may further support physicians' post-exposure performance.

The second question is who benefits. Prior work has examined worker capability and experience in AI collaboration (Wang et al. 2023, Brynjolfsson et al. 2025). We distinguish performance on the common unaided baseline cases from a pre-task self-evaluation of ability relative to colleagues. A one-standard-deviation higher self-evaluation is associated with a 0.287-point smaller cohort change difference during exposure and a 0.338-point smaller difference after withdrawal. The after-withdrawal association remains negative (−0.221) in a joint model with baseline performance and its stage interactions. The findings suggest that being humble might be helpful for instant learning from AI.

Our study makes three contributions to the literature. First, it shifts the focus of AI adoption from productivity gains during assisted work to immediate learning revealed through subsequent unaided performance (Jussupow et al. 2021, Wang et al. 2023, Caplin et al. 2025). The persistence of performance gains on new cases provides evidence of immediate knowledge transfer beyond simple memorization, while the short post-exposure horizon keeps the interpretation focused on rapid capability development rather than long-term learning.

Second, the paper identifies self-evaluation as an important moderator of AI-enabled learning and

empirically distinguishes its role from physicians' observed capability (Noy and Zhang 2023, Brynjolfsson et al. 2025). This finding shows that learning from AI depends not only on what individuals can do, but also on how they perceive their own ability. In doing so, our study introduces self-evaluation as a distinct individual characteristic in the human–AI interaction literature, complementing the existing emphasis on objective capability and experience. The results suggest that individuals with similar levels of expertise may nevertheless learn differently from AI because they differ in how they assess their own competence.

Third, the study extends emerging research on learning from AI beyond student and novice settings to experienced professionals. This distinction is theoretically important, not merely contextual. Students and novices are still developing domain knowledge and problem-solving routines and may therefore be more likely to rely heavily on AI-generated solutions (citations from the education literature). Experienced professionals, by contrast, possess established knowledge structures and judgment frameworks that enable them to evaluate AI-generated evidence, identify which information is informative, and selectively integrate that information into their own decision processes. To our knowledge, relatively few studies of AI-enabled learning have examined professional knowledge workers as the focal study population.

2. Literature Review

2.1 AI Assistance, Productivity and Learning

A growing literature examines how artificial intelligence affects human performance when AI assistance is available. Across writing, software development, customer service, consulting, and medical diagnosis, AI often improves the quality or speed of task completion, although the size of the gains varies across tasks and workers (Noy and Zhang 2023; Peng et al. 2023; Brynjolfsson et al. 2025; Dell'Acqua et al. 2026). Similar patterns appear in healthcare. AI-based decision support can improve diagnostic accuracy and efficiency, but its effects depend on the quality of algorithmic advice and on how clinicians interpret it (Tschandl et al. 2020; Gaube et al. 2021; Yu et al. 2024). These studies establish that AI can augment human judgment at the point of use. However, most evaluate performance while AI remains available and therefore leave open whether interaction with AI changes human capability after the technology is removed.

An emerging stream of research begins to address this distinction between assisted performance and learning. Brynjolfsson et al. (2025) show that less experienced customer-service agents improve disproportionately under AI assistance and move more rapidly along the experience curve, consistent with AI transmitting practices associated with more capable workers. Direct evidence on post-exposure learning, however, has largely come from educational settings. Kestin et al. (2025) find that a carefully designed AI tutor improves student learning, whereas Bastani et al. (2025) show that unrestricted access to generative AI can raise performance during practice but reduce subsequent unaided performance. Together, these studies demonstrate that performing better with AI is not equivalent to learning from AI: AI may either build capability or substitute for the cognitive effort required to develop it.

Our study extends this literature by separating the direct augmentation effect of AI from its post-exposure learning effect in a professional decision environment. Physicians first perform diagnoses without AI, then work with or without AI assistance, and finally complete additional diagnoses after AI is removed. This design allows us to observe whether AI exposure changes subsequent unaided performance rather than merely improving decisions while the technology is present. Moreover, the post-exposure stage includes previously unseen cases. Improvement on these new cases provides evidence of immediate knowledge transfer beyond memory of earlier AI-assisted cases. Accordingly, we view AI not only as a productivity-enhancing tool but also as a potential capability-building technology.

2.2 Heterogeneity in AI Benefits: Capability and Self-Evaluation

A second stream of research documents substantial heterogeneity in the benefits of AI. A recurring finding is that lower-performing or less experienced workers often gain more from AI assistance (Noy and Zhang 2023; Tschandl et al. 2020; Brynjolfsson et al. 2025). Yet objective capability does not fully explain who benefits. In radiology, Yu et al. (2024) find that conventional measures such as experience, subspecialty, and baseline performance do not consistently predict responses to AI, particularly when AI itself is imperfect. These findings suggest that human-AI performance depends not only on what users know, but also on how they interpret and respond to algorithmic information.

Research on algorithm appreciation, reliance, and metacognition points to users' beliefs about their own ability as an additional determinant. Individuals who place greater confidence in their own judgment may rely less on external advice, whereas those who are less certain may be more receptive

to discrepant information (Logg et al. 2019; Chong et al. 2022; Ma et al. 2024). This literature, however, has focused primarily on whether individuals follow AI recommendations while AI advice is available. Much less is known about whether self-evaluation affects how much knowledge users extract from AI and carry into subsequent independent work.

We therefore distinguish objective capability from self-evaluation. Capability captures what a physician can do, whereas self-evaluation captures what a physician believes he or she can do. Two physicians with similar observed diagnostic ability may respond differently to the same AI-generated evidence if they differ in how they assess their own expertise. A physician with lower self-evaluation may scrutinize AI-provided information more carefully and update existing diagnostic heuristics, whereas a physician with higher self-evaluation may perceive less informational value in external input. By examining capability and self-evaluation jointly, our study introduces self-evaluation as a distinct source of heterogeneity in AI-enabled learning rather than treating it as a proxy for skill or experience.

2.3 Learning from AI Among Experienced Professionals

The professional setting is also theoretically important. Much of the direct experimental evidence on AI-enabled learning comes from students and novices who are still developing domain knowledge and problem-solving routines (Kestin et al. 2025, Bastani et al. 2025). For these users, AI may partly substitute for cognitive effort, allowing them to obtain solutions without acquiring the knowledge needed to reproduce those solutions independently. Experienced professionals confront a different learning problem. Through extensive training and repeated practice, they already possess domain-specific knowledge structures, pattern-recognition routines, and judgment frameworks. They can compare AI-generated evidence with established internal models, determine which information is informative, and selectively incorporate useful signals into later decisions. Prior research likewise shows that professionals actively evaluate algorithmic outputs rather than passively accepting them (Lebovitz et al. 2022).

This distinction makes the study of AI learning among experts more than an external-validity extension of student-based research. It raises the possibility that AI can complement an existing judgment framework and accelerate knowledge acquisition rather than replace the reasoning process itself. Using practicing radiologists, our study provides evidence on whether brief exposure to AI-generated diagnostic evidence improves subsequent unaided decisions and whether such learning

transfers immediately to new cases. The study therefore contributes to the human-AI literature in three ways: it shifts attention from productivity during AI use to post-exposure capability, identifies self-evaluation as a moderator distinct from objective capability, and provides evidence on AI-enabled learning among experienced professional knowledge workers.

3. Hypotheses Development

3.1 AI Assistance and Direct Augmentation

AI decision support can improve expert performance by complementing, rather than replacing, professional judgment. In medical imaging, AI can process large volumes of visual information and make subtle diagnostic features more salient (McKinney et al. 2020). Physicians can then combine this algorithmic evidence with their own clinical knowledge and contextual judgment. When the AI system provides structured information about features such as nodule location, size, density, and imaging characteristics, it may reduce the information-processing burden required to identify diagnostically relevant cues and narrow the set of plausible interpretations.

These complementarities can improve both quality and speed. By directing attention to features that might otherwise be overlooked or difficult to quantify, AI assistance can increase diagnostic accuracy (Agrawal et al. 2018, Choudhury et al. 2020). At the same time, by organizing relevant evidence and reducing uncertainty, AI may shorten the time required to reach a diagnostic conclusion. Thus, the immediate value of AI need not arise from substituting for physician judgment; it can arise from helping physicians use their existing expertise more effectively while the technology is present. We refer to this as the direct augmentation effect of AI assistance.

Hypothesis 1: *AI assistance improves physicians' diagnostic performance during use by increasing diagnostic accuracy and reducing decision time.*

3.2 AI Exposure and Post Exposure Learning

The value of AI may extend beyond the period in which the technology is actively used (Gaessler and Piezunka 2023, Choi et al. 2025). Unlike systems that merely provide a final recommendation, the AI system in our setting provides feature-level diagnostic evidence. During AI-assisted diagnosis, physicians can compare this evidence with their own interpretation of the CT image, observe diagnostic cues that they may have previously underweighted, and update the way they map imaging features to diagnostic judgments. Such interaction creates an opportunity for learning because the physician is

exposed not only to an answer but also to information that can be incorporated into an existing diagnostic framework.

This possibility reflects a dynamic form of complementarity between AI and human expertise. Conventional complementarity focuses on whether two inputs generate greater value when used together at the same time. Dynamic complementarity arises when the use of one input increases the future productivity of another. In our setting, AI may temporarily augment diagnosis while also changing how physicians subsequently process images on their own. If exposure helps physicians refine feature recognition, weighting, or diagnostic heuristics, then AI can create value even when it is no longer present. The relevant outcome is therefore not only assisted performance in Stage 2, but also whether unaided performance improves immediately afterward.

If physicians internalize these diagnostic insights, the benefits of AI should persist after assistance is removed. Importantly, true learning should not be confined to previously encountered cases. Memory or task familiarity may improve performance on repeated cases, but transferable learning should also improve decisions on new cases that require physicians to apply the acquired diagnostic knowledge to unfamiliar images. In our setting, the short interval between AI exposure and subsequent unaided decisions allows us to focus on immediate learning rather than long-term skill accumulation. We therefore expect prior AI exposure to generate a post-exposure learning effect that is observable in subsequent independent performance and that can transfer to new cases.

Hypothesis 2. *Exposure to AI assistance improves physicians' subsequent unaided diagnostic performance after AI assistance is removed, including on new cases.*

3.3 Self-Evaluation and Learning from AI

Physicians may learn differently from the same AI-generated evidence. Objective capability matters because lower-performing physicians may have more room to improve, but learning may also depend on self-evaluation—physicians' prior beliefs about their diagnostic ability relative to peers(Yeung and Summerfield 2012, Fleming and Lau 2014). Thus, physicians with similar capability may respond differently to the same AI information.

Physicians with lower self-evaluation may be more receptive to external evidence, compare AI-generated information with their own reasoning, and update their diagnostic judgments. Those with

higher self-evaluation may rely more on established routines and discount conflicting information(Carlebach and Yeung 2023). This distinction is especially relevant for experienced professionals, who already possess established knowledge structures and judgment frameworks.

Accordingly, we expect self-evaluation to shape post-exposure learning beyond objective capability. Capability reflects underlying skill, whereas self-evaluation reflects beliefs about that skill.

Hypothesis 3: *The post-exposure learning effect of AI is greater for physicians with lower self-evaluation.*

4. Experimental Design and Data

4.1 Basic Setup: Case presentations and AI System

We randomly selected 16 case presentations from pathologically confirmed pulmonary nodule cases in a tertiary hospital system. The four target categories form an ordered spectrum: atypical adenomatous hyperplasia (AAH), adenocarcinoma in situ (AIS), minimally invasive adenocarcinoma (MIA), and invasive adenocarcinoma (IAC). Physicians viewed 4–11 curated CT images per case, selected a category, and rated confidence in that decision. The task concerns classification of an already located suspicious nodule; it does not reproduce complete clinical CT interpretation, nodule detection, or treatment decisions. Case allocation and its uneven distribution are described in Supplementary Table S1. For each case, physicians were provided with representative chest CT images along with basic clinical information, including patient age, gender, smoking history, and relevant medical history. All patient identifiers were encrypted to ensure privacy protection.

The AI system used in this study is an intelligent decision-support tool designed for lung nodule analysis based on chest CT imaging. Developed by a leading technology company, it has been integrated into routine clinical practice at the participating hospital system which granted us the access to their CT image cases. The system achieves high accuracy in feature-level decomposition and pattern recognition, enabling it to provide radiologists with rich diagnostic insights beyond simple predictions. The system automatically analyzes CT images and provides structured auxiliary information about lung nodules, including their location, size, density, pleural indentation, and vessel convergence. The system did not provide an AAH/AIS/MIA/IAC classification, category probability, or recommended final answer. In essence, the AI system provides radiologists with diagnostic evidence and subtle imaging features that may be difficult for the human eyes to detect or quantify accurately, while leaving the ultimate

diagnostic judgement to the radiologists.

In the experiment, physicians in the treatment group made diagnostic decisions using the AI-generated outputs in conjunction with the original CT images and patients' basic clinical information. In contrast, physicians in the control group made diagnostic decisions based solely on the CT images and clinical information, without access to AI assistance. Physicians in both groups did not receive pathological truth or correctness feedback after any response. The treatment contrast therefore combines localization and structured imaging features; it cannot isolate a feature, the AI provenance of the information, or visual salience alone. One deployed case had an incorrect density label generated by AI, which was addressed through a case exclusion analysis in Supplementary Table S2.

4.2 Experimental Procedures

The experiment comprised a pre-experimental survey followed by three sequential diagnostic stages. Prior to the task, physicians completed a background questionnaire capturing demographic characteristics, professional background, and self-evaluation.

We randomly allocated the 16 cases across the three experimental stages, with 4 cases assigned to Stage 1, 8 cases to Stage 2, and 4 cases to Stage 3. The four diagnostic categories were unevenly represented in both the broader case population and the 16 cases included in our experiment. For example, only two of the 16 experimental cases were classified as AAH. As a result, radiologists could not infer the diagnoses of subsequent cases based on the categories they had assigned to earlier cases, helping ensure that each diagnosis was made independently based on the available clinical and imaging evidence.

In Stage 1, all physicians diagnosed 4 cases without AI assistance. This stage establishes a baseline measure of individual diagnostic performance in the absence of AI assistance and serves as an objective proxy for physicians' initial diagnostic capability.

In Stage 2, physicians completed diagnoses for 8 new cases under one of two pre-specified conditions. We assigned a cohort of 200 physicians to follow an AI-exposure sequence and complete Stage 2 with AI assistance. To construct a contemporaneous control group and support difference-in-differences estimation, we also assigned an additional 160 physicians to complete Stage 2 without AI assistance. This stage identifies the contemporaneous effect of AI assistance on diagnostic and

operational performance. In the data session, we present the results of a balance check which ensures the similarities in physician characteristics between the treatment and control groups.

In Stage 3, all physicians again diagnosed eight cases without AI assistance. Four of these cases were new cases, while the remaining four were repeats from Stage 1 but presented in a different order. By removing AI assistance in this stage, the design isolates whether prior exposure to AI generates immediate improvements in unaided performance. The inclusion of both repeated and new cases allows us to distinguish learning spillovers from case familiarity. Specifically, this stage tests whether AI decision support exhibits immediate complementarity with human expertise, enhancing not only productivity performance when AI is available, but also subsequent decision quality after the AI assistance is removed.

4.3 Data

The study recruited licensed physicians working in medical imaging departments or related specialties at tertiary hospitals across 153 Chinese cities through a professional online platform. All participants completed the same case presentations in the same order in one session. Fixed compensation of RMB 370 was unrelated to performance. A total of 360 radiologists from tertiary hospitals participated in the study at different time points of our study period. Eligibility required prior experience in interpreting chest CT scans for lung nodules and familiarity with the diagnostic categories of AAH, AIS, MIA, and IAC. Since this study was not a clinical trial, it was not subject to clinical-trial registration and was therefore not publicly preregistered. Data collection took place over a 7-month period in 2025. Supplementary Section S.4 documents give further implementation and ethics information.

To ensure data quality, we applied the following screening criteria. First, we excluded 6 physicians with an average initial diagnostic score below 5, which likely reflects insufficient diagnostic ability or inattentive responding. Second, we excluded 35 physicians who spent less than five seconds on more than two cases, to rule out random or non-deliberative responses. The criteria overlap for two physicians. After applying these criteria, the final sample consisted of 321 physicians, including 134 in the no-AI group and 187 in the AI-assisted group. Table 1 presents summary statistics for variables used in this study.

Given the possibility that differences in participants' arrival times could have affected treatment assignment, we conduct balance checks to evaluate whether the treatment and control groups are comparable on observable characteristics. Table 2 reports balance checks on key physician characteristics across the treatment and control groups. We examine six baseline characteristics: gender, age, education, professional level, nodule diagnostic experience, and self-evaluation. Mean comparisons indicate that these observable characteristics are comparable across groups. Importantly, all empirical specifications include physician fixed effects, so identification relies on within-physician changes over stages, effectively controlling for time-invariant differences.

4.4 Key Variables

4.4.1 Dependent Variables: Diagnostic Accuracy and Decision Time

We focus on two main dependent variables that capture different dimensions of diagnostic performance: diagnostic accuracy, and diagnostic time. Together, these outcomes characterize both quality and speed of physicians' decisions, which are central to evaluating the operational impact of AI-based decision support.

Diagnostic accuracy is the primary performance outcome. It is measured using the graded scoring scheme described earlier and reflects the clinical severity of diagnostic errors. Accuracy captures the core objective of diagnostic decision making and serves as the central metric for evaluating the effectiveness of AI assistance.

In the experiment, physicians were asked to classify each case into one of four lung nodule types (AAH, AIS, MIA, or IAC). Diagnostic performance was quantified using a graded scoring scheme that accounts for the severity of diagnostic errors. A diagnosis that exactly matched the pathological ground truth received 10 points; a diagnosis adjacent to the ground truth in pathological severity (e.g., AIS vs. MIA) received 5 points; all other diagnoses received 0 points. Here, "adjacent" refers to misclassifications that differ by one stage along the pathological progression of lung adenocarcinoma. Accordingly, AAH–AIS, AIS–MIA, and MIA–IAC are defined as adjacent pairs. This definition follows the established classification system in which lung adenocarcinoma progresses from AAH to AIS to MIA to IAC, reflecting increasing invasiveness and worsening prognosis (Travis et al. 2011). This scoring scheme distinguishes different degrees of diagnostic errors and reflects their potential clinical

relevance. As a robustness check, we use an alternative diagnostic performance measure, *exact accuracy*, defined as a dummy variable equal to 1 if the diagnosis is correct and 0 otherwise. This secondary measure avoids the adjacent-category scoring choice.

Decision time measures the number of seconds from display of the case page to submission of classification. It includes viewing and answer entry. The principal time model winsorizes values at the pooled main-sample 1st and 99th percentiles. The platform did not record time specifically viewing the AI panel.

4.4.2 Key Moderating Variable

Our key independent variable is self-evaluation, which captures a physician's prior belief about his or her own ability. *Self-evaluation* is elicited in the pre-experimental survey, where physicians rate their lung nodule diagnostic ability relative to peers on a 1–7 Likert scale, ranging from "far below peers" to "far above peers." This measure reflects physicians' perceived expertise rather than their actual performance. Self-evaluation is central to our analysis because it may shape how physicians engage with AI advice, and whether they internalize AI-provided information into their own decision making. The mean of self-evaluation among all physicians in our final sample are 5.044 out of 7.

To facilitate interpretation across specifications, we standardize self-evaluation using z-scores in all regression analyses. A one-unit change corresponds to a one–standard deviation increase in the underlying construct, with zero representing the sample mean.

5. Empirical Methods

This section outlines the empirical strategy used to examine how AI assistance affects diagnostic performance along three dimensions: accuracy, confidence, and time in different experimental stages. We present both baseline specifications and heterogeneity analyses to characterize the average contemporaneous and immediate learning effects of AI and how these effects vary across physicians with different levels of self-evaluation.

5.1 Direct Augmentation and Post-exposure Learning Effects of AI Assistance

We estimate the contemporaneous effect of AI assistance using a difference-in-differences (DID) specification that exploits within-physician variation across stages, while controlling for physicians,

and case fixed effects. Our baseline empirical specification is:

$$Y_{ict} = \beta_0 + \beta_1 Treat_i \times Stage_t + \alpha_i + \lambda_{c(t)} + \varepsilon_{ict} \quad (1)$$

where Y_{ict} denotes physician i's diagnostic outcome for case c at stage t. The outcome variable captures diagnostic accuracy, or decision time, depending on the specification. $Treat_i$ is an indicator equal to one if physician i is assigned to the AI-assisted condition, and $Stage_t$ equals one for observations in Stage 2 and zero for Stage 1. The coefficient of interest, β_1, estimates the effect of AI assistance on diagnostic performance.

The specification includes physician fixed effects (α_i) to control for unobserved time-invariant individual characteristics, and case-presentation fixed effects (λ_c) to absorb variation shared by physicians assessing the same presentation. The dummy $Treat_i$ is automatically absorbed by the physician fixed effects. Because each case appears in only one stage, a separate stage main effect ($Stage_t$) is nested in the case fixed effects. The error term ε_{ict} captures idiosyncratic shocks that vary across physicians, cases, and stages. The primary standard errors cluster by physician. Supplementary Table S3-S4 reports two-way clustering by physician and underlying case and covariate-adjusted physician-level analyses.

To assess whether exposure to AI assistance generates immediate learning effects beyond contemporaneous use, we re-estimate Equation (1) by comparing diagnostic performance in Stage 3 with baseline performance in Stage 1. In this specification, $Stage_t$ is redefined to equal one for observations in Stage 3 and zero for Stage 1, while all other parameters remain unchanged.

5.2 Heterogenous Effects of AI Assistance by Self-evaluation

Next, we examine heterogeneity in the effect of AI assistance by physicians' self-assessed diagnostic capability, with a particular focus on whether self-evaluation moderates the complementary effect of AI on diagnostic outcomes. This analysis allows us to assess whether AI assistance differentially benefits physicians depending on how they perceive their own expertise.

The heterogeneity specification is given by:

$$Y_{i,c,t} = \beta_0 + \beta_1 Treat_i \times Stage_t + \beta_2 Stage_t \times Selfeval_i + \beta_3 Treat_i \times Stage_t \times Selfeval_i + \alpha_i + \lambda_{c(t)} + \varepsilon_{i,c,t} \quad (2)$$

where all variables are defined as in Equation (1), and $Selfeval_i$ is a standardized continuous measure

of physician i's self-evaluation. Equation (2) extends the baseline specification by allowing the effect of AI assistance to vary systematically with physicians' self-evaluation. Because self-evaluation is a pre-task measure and does not vary over stages, it is orthogonal to AI exposure and is therefore not endogenous to AI usage. The effect of self-evaluation is absorbed by physician fixed effects. The interaction term, $Stage_t \times Selfeval_i$, allows the no-AI cohort's trajectory to vary with self-evaluation. The triple interaction, $Treat_i \times Stage_t \times Selfeval_i$, describes how much the between-cohort change difference changes for one standard deviation higher self-evaluation.

In this specification, the marginal effect of AI assistance on diagnostic outcomes is given by $\beta_1 + \beta_3 \, SelfEval_i$. At the mean level of self-evaluation (z = 0), the marginal effect reduces to β_1. When the outcome is diagnostic accuracy, a positive (negative) estimate of β_3 indicates that physicians with higher self-evaluation experience larger (smaller) performance gains from AI assistance relative to those with lower self-evaluation.

6. Main Results

6.1 Direct Augmentation and Post-exposure Learning Effects of AI Assistance

Table 3 reports the direct augmentation and post-exposure learning effects of AI assistance on diagnostic performance. Columns (1)–(2) report concurrent effects when AI is available in Stage 2. Relative to a baseline mean accuracy of about 6.838 points at stage 1, access to AI raises diagnostic accuracy by roughly 1.399 points, a 20.5% increase. At the same time, AI reduces decision time by 10.983 seconds per case, or a 22% reduction using a log transformation.

Columns (3)–(4) report the immediate learning effects after AI assistance is removed. Physicians previously exposed to AI show an additional 21.6% improvement in diagnostic accuracy from Stage 1 to Stage 3 relative to physicians without AI exposure. They also exhibit about 20% larger reductions in decision time. These patterns suggest that AI exposure generates instant gains in both accuracy and efficiency rather than temporary reliance on the tool.

6.2 Heterogenous Effects of AI Assistance by Self-evaluation

Table 4 examines whether cohort change differences during and after AI exposure vary with physicians' self-evaluation, allowing self-evaluation to predict stage changes in the No-AI control group.

Columns (1)–(2) report Stage 2 results. For diagnostic accuracy score, the $Stage_t \times Selfeval_i$ coefficient is small and statistically insignificant (0.050). The coefficient of the triple interaction $Treat_i \times Stage_t \times Selfeval_i$ is negative (−0.287, $p < 0.05$). Thus, after accounting for the control cohort's self-evaluation-related stage trajectory, higher self-evaluation is associated with a smaller diagnostic accuracy score change difference. The estimated difference is 1.698 points at one standard deviation below mean self-evaluation and 1.124 points at one standard deviation above it.

This pattern is more limited for decision time. The Stage 2 triple interaction for decision time (−1.201) is not statistically significant. Although the estimated cohort change differences indicate shorter time at both low and high self-evaluation, the differences between those conditional estimates are not statistically distinguishable. The observed Stage 2 heterogeneity is therefore concentrated in diagnostic accuracy score.

Columns (3)–(4) examine Stage 3, when neither cohort receives AI-assisted information. The $Stage_t \times Selfeval_i$ coefficient for diagnostic accuracy score is close to zero (0.001), while the triple interaction remains negative (−0.338, $p < 0.05$). The estimated change difference is 1.842 points at low self-evaluation (−1 SD) and 1.166 points at high self-evaluation (+1 SD). Physicians with lower initial self-evaluation therefore exhibit a larger relative increase in their subsequent unaided diagnostic accuracy scores. Stage 3 triple interaction for decision time is not statistically significant. The evidence does not support systematic moderation of diagnostic time.

6.3 Identification and Robustness Checks

6.3.1 Addressing Baseline Differences and an Alternative Skill-Based Explanation

Before turning to the mechanisms underlying the observed learning effects, we address an important identification concern arising from the baseline performance differences between the AI-exposure and control groups. A key limitation of the experimental design is that the treatment and control groups exhibit differences in diagnostic accuracy and completion time in Stage 1. Because participants were recruited sequentially and we do not observe multiple pre-treatment periods, we cannot directly establish parallel pre-treatment trends. Consequently, unobserved differences between the two groups in how their performance might have evolved across subsequent cases remain a potential alternative explanation.

One specific concern is that the observed post-exposure performance improvement could reflect superior underlying ability among physicians assigned to the AI-exposure group rather than the effect of AI itself. If this were the primary mechanism, however, we would expect physicians in the AI-exposure group to exhibit better performance before receiving AI assistance—that is, higher diagnostic accuracy and shorter decision times in Stage 1.

To assess this possibility, we first conduct unadjusted two-sample Welch tests comparing the two groups in Stage 1 (Supplementary Table S5). The results show that, relative to the control group, physicians in the AI-exposure group have diagnostic accuracy that is 0.267 points lower, and decision time that is 8.625 seconds longer. The differences in accuracy and diagnostic time are statistically significant at approximately the 1% level. Importantly, the directions of these two differences are the opposite of what the superior-skill explanation would predict.

These baseline patterns therefore provide little support for the explanation that the estimated performance gains are driven by systematically more capable physicians in the AI-exposure group. If anything, the lower baseline accuracy and longer diagnostic time suggest that the AI-exposure group entered the experiment with weaker observed performance on the two primary dimensions. To the extent that such baseline differences persist and bias the estimates, they would tend to work against finding a positive AI effect rather than mechanically generate one.

Further, as a robustness check, we employ a stricter measure of diagnostic accuracy: *exact accuracy*. Using this measure, the Stage 1 difference between the AI-exposure and control groups is −1.97 percentage points and is statistically insignificant. Thus, physicians in the AI-exposure group do not exhibit higher baseline diagnostic accuracy under this more stringent definition. This result provides additional evidence against the alternative explanation that the subsequent performance gains are driven by superior underlying diagnostic skill among physicians assigned to the AI-exposure group.

In addition, we exclude physicians whose Stage 1 mean accuracy falls below 0.25, corresponding to chance performance in a four-category classification task. Table 5 shows that the magnitude remains consistent. Nevertheless, these comparison and robustness checks cannot rule out all forms of unobserved heterogeneity or differential performance trends between the two groups. We therefore interpret the results with this limitation in mind and supplement the main analysis with controls and robustness analyses designed to account for observable baseline differences.

6.3.2 Other Robustness Checks

We conduct two sets of additional robustness checks, reported in Supplementary Table S6-7. First, we are concerned that the results might be sensitive to different sample screening criteria. Table S6 reports the results using different samples. The results remain qualitatively and qualitatively similar. Second, we concern about the impact of AI hallucination. To alleviate the concern, we restrict the sample to physicians in the AI-exposed group and estimate within-physician changes across stages, explicitly controlling for AI errors. Here, *AI_error* is an indicator that equals one when the AI system provides incorrect auxiliary information about nodule characteristics (e.g., density) for a given case and zero otherwise. Because this robustness check focuses only on physicians who were exposed to AI, the treatment indicator is no longer included. We therefore estimate a within-physician specification with physician fixed effects and include *AI_error* as an additional control variable (see Table S7). The resulting estimates are again highly consistent with our baseline results, indicating that our conclusions are not driven by the choice of ability measure, the functional form of the accuracy outcome, sample screening, or unmodeled AI errors.

7. Mechanism and Extension

7.1 Instant Transfer versus Task Familiarity

A potential concern is that post-exposure performance improvements reflect task familiarity or memorization rather than learning from AI. Our design mitigates this concern in two ways. First, both treatment and control physicians complete the same multi-stage diagnostic task, so generic familiarity, repetition, and memory effects are absorbed by the control group and differenced out in our empirical comparisons. Second, Stage 3 includes both repeated and new cases. If the immediate transfer effects are driven mainly by memorization, the gains should be concentrated on repeated cases. If they reflect transferable learning, similar gains should appear on new cases.

Table 6a separates new and repeated cases in Stage 3 to examine whether the post-exposure pattern in diagnostic accuracy score is primarily attributable to memory of previously seen cases. Columns (1) and (3) show that prior AI exposure significantly improves accuracy in both settings. The effect is 1.646 points for new cases and 1.311 points for repeated cases, corresponding to 24.1% and 19.2% of the

baseline mean accuracy at stage 1, respectively. Because the improvement is not confined to repeated cases, the results provide little support for a pure memorization explanation. Instead, they suggest that the post-exposure performance improvement extends beyond repeated cases to new diagnostic tasks.

Columns (2) and (4) further examine whether these post-exposure change differences vary with physicians' self-evaluation. For new cases, the coefficient on $Stage_t \times Selfeval_i$ is small and insignificant (0.077), whereas the coefficient on $Treat_i \times Stage_t \times Selfeval_i$ is negative and marginally significant (−0.365, $p < 0.10$). For repeated cases, the corresponding stage interaction is also insignificant (−0.075), while the triple interaction is negative and marginally significant (−0.311, $p < 0.10$). The negative triple-interaction estimates in both case types suggest that lower self-evaluation may be associated with larger post-exposure diagnostic accuracy score change differences, after allowing self-evaluation to predict stage-related changes in the No-AI comparison cohort. However, the heterogeneity evidence should be interpreted cautiously because the triple interactions are only marginally significant in both subsamples. Accordingly, Table 6a provides clearer evidence that the average post-exposure pattern extends to new cases than that self-evaluation moderates this pattern within each case type.

Table 6b directly tests the difference between the two average Stage 3 estimates in a joint model containing both case types. The new-case estimate exceeds the repeated-case estimate by 0.334 points, but this difference is not significant (SE = 0.238, 95% CI [−0.135, 0.803], $p = 0.162$). The data therefore show a positive score change difference on previously unseen cases without establishing a larger difference for new than repeated cases.

7.2 Greater Effort or Attention During Later Stages

Another potential concern is that physicians in the AI-assisted group may have exerted greater effort or paid more attention to the CT images, leading to improved performance in Stage 3. Under this interpretation, the observed improvement would reflect increased effort rather than learning from AI exposure.

While this is a plausible alternative explanation, it is difficult to reconcile with the pattern of our results. If the performance improvements were driven primarily by increased effort or attention, one would expect physicians to spend more time evaluating each case as they allocate additional cognitive

resources to the diagnostic task. Instead, we observe the opposite pattern: diagnostic accuracy improves while decision time declines. This combination of higher accuracy and shorter decision time suggests that physicians process diagnostic information more efficiently after AI exposure, rather than simply working harder or spending more time on the task.

The heterogeneity results provide additional evidence against a pure effort-based explanation. Physicians with lower self-evaluations experience larger improvements in diagnostic accuracy, yet their decision times are comparable to those of physicians with higher self-evaluations. Thus, the greater performance gains among lower-self-evaluation physicians do not appear to arise from greater time investment. Taken together, these patterns are more consistent with an improvement in diagnostic capability—potentially resulting from learning during AI-assisted diagnosis—than with an explanation based solely on increased effort or attention.

7.3 Diagnostic Confidence as a Potential Mechanism

One potential mechanism underlying the post-exposure performance improvement in Stage 3 is increased diagnostic confidence. Exposure to AI may strengthen physicians' confidence in their diagnostic judgments, which in turn may improve their subsequent performance after AI assistance is removed. Diagnostic confidence is an important dimension of clinical decision making because it reflects physicians' perceived competence and may shape how they interpret diagnostic information and make subsequent decisions. Examining confidence alongside objective diagnostic accuracy therefore allows us to assess whether AI exposure affects physicians' perceived competence in addition to their observed performance.

To examine this potential mechanism, we use *diagnostic confidence*, a case-level measure of physicians' certainty in their diagnoses. After completing the diagnosis for each case, physicians reported their confidence on a 1–7 scale.

Table 7 reports the results. Columns (1) and (2) first examine whether AI exposure affects diagnostic confidence during and after AI assistance. Column (1) shows that physicians receiving AI assistance in Stage 2 report confidence that is 0.286 points higher, corresponding to a 5.7% increase relative to the Stage 1 baseline of 5.032. More importantly, the increase in confidence persists after AI assistance is removed. As shown in Column (2), physicians previously exposed to AI exhibit

approximately 3.2% higher diagnostic confidence in Stage 3, when both groups perform the diagnostic task without AI support. These findings suggest that AI exposure not only increases physicians' confidence while they are using the technology but also has a spillover effect on their perceived diagnostic competence after AI assistance is withdrawn.

Columns (3) and (4) then examine whether increased confidence may account for part of the post-exposure improvement in diagnostic accuracy and time observed in Stage 3. The specification for the medication analysis is as follows:

$$Y_{ict} = \beta_0 + \beta_1 Treat_i \times Stage_t + \gamma ConfidenceProxy_{it} + \alpha_i + \lambda_{c(t)} + \varepsilon_{ict} \quad (3)$$

Importantly, we use confidence measures that precede the focal diagnosis, thereby reducing concerns that the relationship simply reflects physicians becoming more confident after arriving at a correct diagnosis. In Column (3) and (4), we construct a prior-confidence measure based on physicians' average confidence in the preceding stage: for Stage 1 observations, because physicians had not yet been exposed to AI, we assume that their underlying confidence remained unchanged and use each physician's average confidence in Stage 1 as the prior-confidence measure. For Stage 3 observations, we use the physician's average confidence in Stage 2.

The two specifications yield a consistent pattern. First, the confidence measure is positively and significantly associated with diagnostic accuracy and negatively and significantly associated with diagnostic time. Second, after this confidence measure is included in the model, the estimated post-exposure effect of AI on diagnostic accuracy remains positive and significant, and becomes slightly smaller than in the corresponding specification without the confidence measure. Similarly, the estimated reduction in diagnostic time remains significant and becomes slightly smaller in magnitude. Taken together, these findings are consistent with diagnostic confidence serving as one channel through which AI exposure improves physicians' subsequent performance. However, its contribution is relatively marginal.

7.4 True Capability or Overestimate of Own Ability

Session 6 shows that AI assistance improves diagnostic performance both when it is available and after it is removed, and that these gains are larger for physicians with lower self-evaluation. A central question is what self-evaluation captures. It may proxy for aptitude, because lower-ability

physicians have more room to improve. Alternatively, it may reflect attitude, insofar as physicians who overestimate their own ability may be less receptive to learning from AI. To distinguish these possibilities, we therefore bring capability and self-evaluation into the same empirical framework, and we further examine whether the moderating role of self-evaluation persists within capability-based subsamples.

We define *Capability*, a performance-based measure, as each physician's mean diagnostic score in Stage 1, when all cases are diagnosed without AI. It summarizes baseline diagnostic skill in the absence of AI and is distinct from self-evaluation. This measure is collected prior to random assignment and is therefore not influenced by AI exposure. The joint specification is as follows:

$$\begin{aligned} Y_{i,c,t} = \beta_0 + \beta_1 Treat_i \times Stage_t + \beta_2 Stage_t \times Capacity_i \\ + \beta_3 Treat_i \times Stage_t \times Capacity_i + \beta_4 Stage_t \times Selfeval_i \\ + \beta_5 Treat_i \times Stage_t \times Selfeval_i + \alpha_i + \lambda_{c(t)} + \varepsilon_{i,c,t} \end{aligned} \quad (4)$$

7.4.1 Direct Augmentation Effect by Capability and Self-evaluation

Table 8 examines baseline capability and self-evaluation jointly in relation to Stage 2 diagnostic accuracy score change differences. In columns (1) and (2), $Stage_t \times Capability_i$ coefficients are −0.845 and −0.844 and highly significant, while the triple interactions are −0.141 and −0.129 and insignificant. Lower baseline performance is therefore associated with greater Stage 1-to-Stage 2 improvement in the comparison cohort, without evidence of a reliably different gradient in the AI-exposure cohort. Greater room for improvement is one possible explanation.

Column (2) adds self-evaluation. Its triple interaction is −0.173 and marginally significant, whereas the $Stage_t \times Selfeval_i$ term is small and insignificant. Higher self-evaluation is thus associated with a smaller Stage 2 score change difference conditional on measured baseline capability and its stage interactions. Specifically, physicians with higher self-evaluation experience smaller gains from AI assistance than those with lower self-evaluation, even after holding capability constant.

Columns (3)–(6) report subgroup estimates. The self-evaluation triple interaction is −0.322 in the lower-capability subgroup (p = 0.035) and −0.063 in the higher-capability subgroup (p = 0.647). The association is statistically more evident in the lower-capability subgroup, but the estimated slopes do not differ significantly across capability groups.

Table 8 therefore shows different associations for baseline capability and self-evaluation. Baseline capability is primarily related to changes shared with the comparison cohort, whereas self-evaluation relates to the additional score change difference. One possible interpretation is that physicians with more modest assessments of their expertise adjust their judgments more in response to structured cues. This remains an interpretation rather than an observed process. Self-evaluation complements the task-specific capability measure in describing heterogeneity in the cohort change differences.

7.4.2 Post-Exposure Learning Effect by Capability and Self-evaluation

Table 9 turns to Stage 3, after AI-assisted information has been withdrawn. It examines whether the score advantage associated with prior AI exposure appears in immediate unaided judgments and how that change difference varies with baseline capability and self-evaluation.

The $Stage_t \times Capability_i$ coefficients in columns (1) and (2) are −0.886 and −0.887 and highly significant. The corresponding triple interactions are −0.127 and −0.104 and insignificant. Lower baseline capability is associated with greater Stage 1-to-Stage 3 improvement, but this gradient is not statistically specific to the AI-exposure cohort. Table 9 therefore does not support a claim that lower-capability physicians have greater AI-specific post-exposure gains.

In the joint model, the $Stage_t \times Selfeval_i$ coefficient is small and insignificant (−0.031), while its cohort triple interaction is negative (−0.221, $p < 0.05$). Higher self-evaluation is associated with a smaller Stage 3 score change difference after accounting for measured baseline capability and its stage-related terms.

The subgroup models in columns (3)–(6) yield self-evaluation triple interactions of −0.384 among lower-capability physicians ($p = 0.005$) and −0.080 among higher-capability physicians ($p = 0.601$). The direct difference is insignificant: −0.304, SE = 0.205, 95% CI [−0.707, 0.099], $p = 0.138$. The association is estimated more clearly in the lower-capability group, but these results do not establish two statistically distinct capability regimes.

Taken together, Table 9 distinguishes general improvement across stages from heterogeneity in the short-run gains associated with prior AI exposure. Baseline capability is related primarily to the improvement shared by the treatment and control groups, whereas self-evaluation differentiates the additional Stage 3 improvement associated with prior AI exposure. This negative relationship is

estimated most clearly among lower-capability physicians, although the interaction coefficients do not differ significantly between the two capability groups.

The overall pattern is consistent with a near-term recalibration of unaided judgment: exposure to structured AI information may temporarily alter how physicians weigh diagnostic cues, with physicians holding more modest assessments of their expertise carrying more of this adjustment into subsequent decisions after AI assistance is removed. Moreover, because Stage 3 followed shortly after AI exposure within the same experimental session, the findings concern the near-term carryover of AI-related accuracy gains into unaided performance. Whether these gains persist over longer periods remains an important question for future research.

8. Discussion and Conclusion

The study documents two related patterns in a short expert decision task. AI-generated localization and imaging features accompany higher accuracy and shorter time while presented. More unusually, the exposure cohort also shows higher accuracy on the next unaided cases, including previously unseen cases. We label this an immediate learning pattern because it involves performance after assistance has ended and extends beyond recalling a repeated image. The task gives no outcome feedback, so any learning would involve interpreting or reweighting the supplied imaging cues rather than verification against the pathological truth.

Prior research shows mixed conclusions regarding whether algorithmic support fosters or crowds out learning (e.g., Risko and Gilbert 2016; Chassagnon et al. 2023; Sundaresan and Guler 2025; Gonzalez 2005; Brynjolfsson et al. 2025). Our evidence contributes to this debate by leveraging the repeated-versus-new case design. We find that post-exposure learning effects in accuracy and efficiency are comparable across repeated and new case. These patterns suggest that physicians instantly internalize elements of AI-supported reasoning and transfer them to unfamiliar cases.

The self-evaluation association adds a worker dimension to this operational question. Measured baseline performance describes skill on four cases; self-evaluation describes the physician's perceived standing relative to colleagues before exposure. The two need not coincide. The negative self-evaluation interaction survives adjustment for the baseline performance trajectory in the immediate unaided stage. This makes beliefs worth measuring when studying who benefits from AI-supported work.

Overconfident physicians appear less responsive to AI input and internalize less from exposure. However, it does not imply that lower self-evaluation should be induced or that more confident experts should be denied access to AI.

These findings have implications for AI deployment in clinical settings. First, the results highlight the importance of metacognitive calibration: with AI, aptitude shapes what you can do; attitude shapes how much more you can learn. Second, evaluations of clinical AI should incorporate dynamic outcomes, including its function as a learning technology and post-exposure trajectories, rather than focusing solely on contemporaneous accuracy.

More broadly, AI assistance should be viewed as a learning technology that provides information physicians may otherwise overlook, rather than merely delivering final judgments. Our findings speak to professional decision environments that require complex visual pattern recognition, such as pathology, dermatology, industrial quality inspection, security screening, and remote-sensing interpretation. In these settings, experts often need to integrate multiple visual cues under uncertainty, and AI assistance systems increasingly provide structured signals that can guide expert judgment. Our results suggest that such tools should not be evaluated only by whether they improve immediate productivity. They should also be evaluated by whether they help experts develop transferable decision skills.

For healthcare organizations adopting AI, the main implication is that workflow design matters. If users passively accept algorithmic recommendations, AI may improve short-run output without strengthening human expertise. In contrast, workflows that encourage users to incorporate AI output into their own judgments are more likely to turn AI into a learning tool. This suggests that organizations should pair AI adoption with diagnostic evidence and subtle feature that promote active engagement, calibrated self-assessment, and knowledge transfer.

Several limits shape the interpretation. First, only 16 selected nodules underlie the 20 presentations, and Stage 2 does not include AAH; inference that accounts for shared cases is less precise, though the average accuracy pattern remains positive. Second, the observed differences between the two cohorts in diagnostic accuracy and decision time at Stage 1 may reflect unobserved physician characteristics that could also influence the magnitude of the effect of AI assistance. Third, Stage 3 occurred immediately after Stage 2 in one session. We cannot infer retained learning, improved clinical outcomes,

or a long-term return to AI adoption.

The operational implication is a measurement principle: evaluation of an AI tool should include subsequent unaided tasks when worker learning matters. Future work could randomize AI access within simultaneous recruitment cohorts, vary cue presentation or reflection opportunities, measure attention to AI features, and retest workers days or weeks later. Such designs could determine whether the immediate carryover seen here becomes durable capability and whether self-evaluation is a causal driver of learning.

References

Agrawal A, Gans J, Goldfarb A (2018) Prediction, judgment, and complexity: A theory of decision-making and artificial intelligence. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 89–110.

Ahmadzade M, Rouientan H, Abdi N, Norouzi M, Hakimi M, Bahrambeigi M, Khalili Pouya E, et al. (2026) AI assistance improves reader agreement in digital mammography: A multireader crossover study of general and breast subspecialty radiologists. *Sci. Rep.* 16(1):1370.

Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *Quart. J. Econom.* 140(2):889–942.

Budzyń K, Romańczyk M, Kitala D, Kołodziej P, Bugajski M, Adami HO, Blom J, et al. (2025) Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: A multicentre, observational study. *Lancet Gastroenterol. Hepatol.* 10(10):896–903.

Caplin A, Deming D, Li S, Martin D, Marx P, Weidmann B, Ye KJ (2025) The ABCs of who benefits from working with AI: Ability, beliefs, and calibration. *Management Sci.*, ePub ahead of print October 24, https://doi.org/10.1287/mnsc.2024.08994.

Carlebach N, Yeung N (2023) Flexible use of confidence to guide advice requests. *Cognition* 230:105264.

Chassagnon G, Billet N, Rutten C, Toussaint T, Cassius de Linval Q, Collin M, Lemouchi L, et al. (2023) Learning from the machine: AI assistance is not an effective learning tool for resident education in chest x-ray interpretation. *Eur. Radiol.* 33(11):8241–8250.

Cheng C-T, Chen C-C, Fu C-Y, Chaou C-H, Wu Y-T, Hsu C-P, Chang C-C, et al. (2020) Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights Imaging* 11(1):119.

Choi S, Kang H, Kim N, Kim J (2025) How does artificial intelligence improve human decision-making? Evidence from the AI-powered Go program. *Strategic Management J.* 46(6):1523–1554.

Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management J.* 41(8):1381–1411.

Dargnies M-P, Hakimov R, Kübler D (2024) Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Sci.* 72(1):285–301.

Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L (2006) Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. *J. Amer. Medical Assoc.* 296(9):1094–1102.

De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, et al. (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* 24(9):1342–1350.

Diekhoff T, Kainberger F, Oleaga L, Dewey M, Zimmermann E (2020) Effectiveness of the clinical decision support tool ESR eGUIDE for teaching medical students the appropriate selection of imaging tests: Randomized cross-over evaluation. *Eur. Radiol.* 30(10):5684–5689.

Dunning D (2011) The Dunning–Kruger effect: On being ignorant of one's own ignorance. Olson JM, Zanna MP, eds. *Advances in Experimental Social Psychology*, vol. 44 (Academic Press, Cambridge, MA), 247–296.

Durieux P, Nizard R, Ravaud P, Mounier N, Lepage E (2000) A clinical decision support system for prevention of venous thromboembolism: Effect on physician behavior. *J. Amer. Medical Assoc.* 283(21):2816–2821.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118.

Fleming SM, Lau HC (2014) How to measure metacognition. *Front. Human Neurosci.* 8:443.

Gaessler F, Piezunka H (2023) Training with AI: Evidence from chess computers. *Strategic Management J.* 44(11):2724–2750.

Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M (2021) Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digit. Med.* 4(1):31.

Gokkurt Yilmaz BN, Ozbey F, Yilmaz BE (2025) Effect of artificial intelligence-assisted personalized feedback on radiographic diagnostic performance of dental students: A controlled study. *BMC Med. Educ.* 25(1):1403.

Gonzalez C (2005) Decision support for real-time, dynamic decision-making tasks. *Organ. Behav. Human Decision Processes* 96:142–154.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, et al. (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *J. Amer. Medical Assoc.* 316(22):2402–2410.

Huang H-Y, Huang Y-H, Lin C-H, Tao W-T, Liao W-C, Yu S, Mo H-C, et al. (2026) AI-assisted chest radiograph interpretation enhances diagnostic confidence and standardizes diagnostic accuracy across radiologists: A multi-reader study. *Clin. Imaging* 130:110694.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary Task Ordering: Queue Management in Radiological Services. Management Science 64(9):4389-4407.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Inform. Systems Res.* 32(3):713–735.

Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, Yu Y, et al. (2020) Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digit. Med.* 3(1):23.

Koh FH, Ladlad J, Foo F-J, Tan WJ, Sivarajah SS, Ho LML, Ng J-L, et al. (2023) Real-time artificial intelligence (AI)-aided endoscopy improves adenoma detection rates even in experienced endoscopists: A cohort study in Singapore. *Surg. Endosc.* 37(1):165–171.

Krakowski I, Kim J, Cai ZR, Daneshjou R, Lapins J, Eriksson H, Lykou A, Linos E (2024) Human-AI interaction in skin cancer diagnosis: A systematic review and meta-analysis. *npj Digit. Med.* 7(1):78.

Küper A, Lodde GC, Livingstone E, Schadendorf D, Krämer N (2025) Psychological factors

influencing appropriate reliance on AI-enabled clinical decision support systems: Experimental web-based study among dermatologists. *J. Med. Internet Res.* 27:e58660.

Liao C, Wen X, Li S, Du P (2024) How effective is AI augmentation in human–AI collaboration? Evidence from a field experiment. *Inform. Tech. People* 37(7):2357–2389.

McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577(7788):89–94.

Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187–192.

Okumura T, Kudo SE, Ide Y, Kato S, Miyata Y, Takisima K, Takashina Y, et al. (2026) Long-term impact of computer-aided adenoma detection: A prospective observational study. *Endoscopy* 58(2):121–129.

Risko EF, Gilbert SJ (2016) Cognitive offloading. *Trends Cogn. Sci.* 20(9):676–688.

Sundaresan S, Guler I (2025) Algorithmic recommendation tools and experiential learning in clinical care. *Organ. Sci.* 36(5):1786–1802.

Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, Beer DG, et al. (2011) International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Oncol.* 6(2):244–285.

Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, Janda M, et al. (2020) Human-computer collaboration for skin cancer recognition. *Nat. Med.* 26(8):1229–1234.

Vasey B, Ursprung S, Beddoe B, Taylor EH, Marlow N, Bilbro N, Watkinson P, McCulloch P (2021) Association of clinician diagnostic performance with machine learning-based decision support systems: A systematic review. *JAMA Network Open* 4(3):e211276.

Vicente L, Matute H, Fregosi C, Cabitza F (2025) Machine learning systems as mentors in human learning: A user study on machine bias transmission in medical training. *Internat. J. Human-Computer Stud.* 198:103474.

Wang W, Gao G, Agarwal R (2023) Friend or foe? Teaming between artificial intelligence and workers with variation in experience. *Management Sci.* 70(9):5753–5775.

Wu S, Liu Y, Ruan M, Chen S, Xie X-Y (2025) Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. *Sci. Rep.* 15(1):15105.

Yeung N, Summerfield C (2012) Metacognition in human decision-making: Confidence and error monitoring. *Philos. Trans. Roy. Soc. B Biol. Sci.* 367(1594):1310–1321.

Yin J, Ngiam KY, Tan SS-L, Teo HH (2025) Designing AI-based work processes: How the timing of AI advice affects diagnostic decision making. *Management Sci.* 71(11):9361–9383.

Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P (2024) Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* 30(3):837–849.

Appendix

A.1 Tables

(1) Summary Statistics and Balance Check

Table 1. Summary Statistics

Variable	N	Mean	SD	Definition
Panel A. Diagnostic outcomes and moderators				
Diagnostic Accuracy	321	7.127	0.820	Physician mean graded diagnostic score across 20 presentations; each diagnosis receives 0, 5, or 10.
Diagnostic Time	321	27.931	10.569	Physician mean seconds per presentation, using pooled main-sample 1st/99th percentile winsorization in the principal time models.
Diagnostic Confidence	321	5.235	0.675	Physician mean self-reported confidence across 20 presentations; 1–7.
Capability	321	6.682	1.154	Mean Diagnostic Accuracy score over the four unaided Stage 1 presentations; standardized in the main capability models.
Self-evaluation	321	5.044	1.201	Pre-task rating of diagnostic ability relative to peers; 1–7, standardized in regressions as $Selfeval_i$.
Panel B. Baseline physician characteristics				
Gender	321	0.371	0.484	Female = 1; male = 0.
Age	321	39.377	9.422	Physician age in years.
Education	321	4.657	1.370	Pre-task highest education category. Full frequencies are in Table 2;
Professional Title	321	2.271	1.021	Pre-task professional-title category. Full frequencies are in Table 2;
Nodule Diagnosis Experience	321	12.128	8.100	Years of experience diagnosing lung nodules before the task.

Notes. N = 321 physicians. Summary statistics use one physician-level record and the original measurement scale; standardized versions are used where specified in the models. SD is the sample standard deviation.

Table 2: Balance Check

Characteristic	AI exposure (n = 187)	No-AI comparison (n = 134)	SMD	p-value
Gender: female, n (%)	71 (38.0%)	48 (35.8%)	0.044	0.726

Characteristic	AI exposure (n = 187)	No-AI comparison (n = 134)	SMD	p-value
Age (years)	39.19 (9.77)	39.64 (8.95)	−0.049	0.666
Education, n (%)				0.484
Associate degree	2 (1.1%)	0 (0.0%)	0.147	
Current undergraduate student	2 (1.1%)	1 (0.7%)	0.034	
Bachelor's degree	50 (26.7%)	38 (28.4%)	−0.036	
Current master's student	9 (4.8%)	14 (10.4%)	−0.213	
Master's degree	91 (48.7%)	57 (42.5%)	0.123	
Current doctoral student	5 (2.7%)	3 (2.2%)	0.028	
Doctoral degree	28 (15.0%)	21 (15.7%)	−0.019	
Professional Title, n (%)				0.874
Junior	48 (25.7%)	37 (27.6%)	−0.044	
Intermediate	68 (36.4%)	46 (34.3%)	0.042	
Associate senior	40 (21.4%)	32 (23.9%)	−0.059	
Senior	31 (16.6%)	19 (14.2%)	0.066	
Nodule Diagnosis Experience (years)	11.89 (8.28)	12.46 (7.86)	−0.070	0.537
Self-evaluation (1–7)	5.13 (1.16)	4.92 (1.24)	0.179	0.117

Notes. N = 321; categorical variables are n (%), other variables means (sample SDs). SMD = (AI mean − No-AI mean) / √[(AI sample variance + No-AI sample variance)/2]; category SMDs use binary indicators. Fisher exact tests compare Gender, Education, and Professional Title, with one overall p-value for each categorical characteristic; other p-values are two-sided Welch tests.

Table 3. Average performance changes during and after AI access

Term	(1) Stage 2 Accuracy Score	(2) Stage 2 Time	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Treat × Stage	1.399*** (0.148)	−10.983*** (1.887)	1.479*** (0.147)	−10.734*** (1.837)
Constant	6.443*** (0.058)	35.664*** (0.733)	6.545*** (0.057)	31.915*** (0.714)
Physician Fixed Effects	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	Yes	Yes
Stage Main Effect	Absorbed	Absorbed	Absorbed	Absorbed

Term	(1) Stage 2 Accuracy Score	(2) Stage 2 Time	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Observations	3852	3852	3852	3852
Physicians	321	321	321	321

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. Score uses the 0/5/10 scale; time is in winsorized seconds. Stage main effects are nested in presentation effects.

(3) Heterogenous Effects of AI Assistance by Self-evaluation

Table 4. Heterogeneity by self evaluation

Term	(1) Stage 2 Accuracy Score	(2) Stage 2 Time	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Treat × Stage	1.411***	−10.795***	1.504***	−10.527***
	(0.148)	(1.895)	(0.148)	(1.853)
Stage × Self-evaluation	0.050	−0.548	0.001	−0.861
	(0.088)	(0.876)	(0.097)	(1.044)
Treat × Stage × Self-evaluation	−0.287**	−1.201	−0.338**	−0.698
	(0.133)	(1.816)	(0.132)	(1.763)
At self-evaluation −1 SD	1.698***	−9.594***	1.842***	−9.829***
	(0.199)	(2.691)	(0.187)	(2.582)
At self-evaluation +1 SD	1.124***	−11.996***	1.166***	−11.225***
	(0.199)	(2.556)	(0.209)	(2.533)
Constant	6.446***	35.626***	6.546***	31.855***
	(0.058)	(0.735)	(0.058)	(0.720)
Physician Fixed Effects	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	Yes	Yes
Stage Main Effect	Absorbed	Absorbed	Absorbed	Absorbed
Observations	3852	3852	3852	3852
Physicians	321	321	321	321

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. Self-evaluation is standardized. All models include physician and presentation effects; stage main effects are nested.

Table 5. Average score gains with a baseline exact accuracy screen

Term	(1) Stage 2 Accuracy Score	(2) Stage 3 Accuracy Score
Treat × Stage	1.377*** (0.149)	1.452*** (0.148)
Constant	6.467*** (0.057)	6.570*** (0.057)
Physician Fixed Effects	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes
Stage Main Effect	Absorbed	Absorbed
Observations	3780	3780
Physicians	315	315

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. Two result columns report only the average effects. Baseline exact accuracy must be at least 0.25 in addition to the main-sample rules. Six additional physicians are excluded.

Table 6a. New and repeated case accuracy score changes in Stage 3

Term	(1) New Average	(2) New Moderation	(3) Repeated Average	(4) Repeated Moderation
Treat × Stage 3	1.646*** (0.201)	1.659*** (0.201)	1.311*** (0.177)	1.348*** (0.177)
Stage 3 × Self-evaluation	—	0.077 (0.153)	—	−0.075 (0.116)
Treat × Stage 3 × Self-evaluation	—	−0.365* (0.188)	—	−0.311* (0.162)
Separate Stage 3 dummy	Nested	Nested	Yes	Yes
Stage 3 dummy	Absorbed	Absorbed	−0.028 (0.138)	−0.036 (0.140)
Constant	6.491*** (0.059)	6.495*** (0.059)	6.682*** (0.043)	6.682*** (0.042)
Physician Fixed Effects	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	No	No
Underlying-case Fixed Effects	No	No	Yes	Yes

Term	(1) New Average	(2) New Moderation	(3) Repeated Average	(4) Repeated Moderation
Observations	2568	2568	2568	2568
Physicians	321	321	321	321

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. Each model includes four Stage 1 cases and four Stage 3 cases. New-case models use physician and presentation effects. Repeated-case models use physician and underlying-case effects plus a separate Stage 3 dummy.

Table 6b. Direct comparison of new and repeated case gains

Contrast	Estimate (SE)	95% CI	p-value
New cases	1.646*** (0.202)	[1.249, 2.042]	<0.001
Repeated cases	1.311*** (0.177)	[0.963, 1.659]	<0.001
New minus repeated	0.334 (0.238)	[-0.135, 0.803]	0.162

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. One joint model includes 321 physicians and 3,852 observations, with physician and presentation effects.

Table 7. Diagnostic confidence and exploratory accuracy associations

Term	(1) Stage 2 Confidence	(2) Stage 3 Confidence	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Treat × Stage	0.286*** (0.053)	0.163*** (0.054)	1.314*** (0.153)	−9.215*** (1.860)
ConfidenceProxy	—	—	0.577*** (0.165)	−5.305*** (2.008)
Constant	5.122*** (0.020)	5.135*** (0.021)	3.592*** (0.848)	59.091*** (10.378)
Physician Fixed Effects	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	Yes	Yes
Stage Main Effect	Absorbed	Absorbed	Absorbed	Absorbed
Observations	3852	3852	3852	3852

Term	(1) Stage 2 Confidence	(2) Stage 3 Confidence	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Physicians	321	321	321	321

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. All models include physician and presentation effects. Column (3) uses the preceding case in the full sequence, omitting each physician's first case. Column (4) uses Stage 1 mean confidence for baseline and Stage 2 mean for Stage 3. The covariate is not randomized. The Stage 1 mean in column (4) is not temporally prior to every Stage 1 diagnosis; this is an exploratory association rather than a causal mediation test.

Table 8. Stage 2 accuracy score changes by capability and self evaluation

Term	(1) All Capability	(2) All Joint	(3) Low Average	(4) Low Self-eval	(5) High Average	(6) High Self-eval
Treat × Stage	1.190*** (0.105)	1.201*** (0.104)	1.907*** (0.176)	1.901*** (0.175)	0.908*** (0.177)	0.930*** (0.175)
Stage × Capability	−0.845*** (0.084)	−0.844*** (0.084)	—	—	—	—
Treat × Stage × Capability	−0.141 (0.120)	−0.129 (0.119)	—	—	—	—
Stage × Self-evaluation	—	0.019 (0.070)	—	0.096 (0.114)	—	−0.045 (0.112)
Treat × Stage × Self-evaluation	—	−0.173* (0.097)	—	−0.322** (0.152)	—	−0.063 (0.137)
Constant	6.519*** (0.041)	6.520*** (0.041)	5.742*** (0.068)	5.746*** (0.068)	7.122*** (0.069)	7.117*** (0.069)
Physician Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Stage Main Effect	Absorbed	Absorbed	Absorbed	Absorbed	Absorbed	Absorbed
Observations	3852	3852	1896	1896	1956	1956
Physicians	321	321	158	158	163	163

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. All models include physician and presentation effects. Low capability is a Stage 1 mean score below 7.5; high capability is at least 7.5. Both continuous moderators are standardized in the full main sample.

Table 9. Stage 3 accuracy score changes by capability and self evaluation

Term	(1) All Capability	(2) All Joint	(3) Low Average	(4) Low Self-eval	(5) High Average	(6) High Self-eval
Treat × Stage	1.261*** (0.100)	1.286*** (0.100)	1.782*** (0.157)	1.778*** (0.154)	1.186*** (0.170)	1.237*** (0.175)
Stage × Capability	−0.886*** (0.081)	−0.887*** (0.081)	—	—	—	—
Treat × Stage × Capability	−0.127 (0.106)	−0.104 (0.104)	—	—	—	—
Stage × Self-evaluation	—	−0.031 (0.073)	—	0.064 (0.105)	—	−0.132 (0.136)
Treat × Stage × Self-evaluation	—	−0.221** (0.092)	—	−0.384*** (0.137)	—	−0.080 (0.153)
Constant	6.625*** (0.039)	6.623*** (0.039)	5.988*** (0.061)	5.991*** (0.060)	7.085*** (0.066)	7.072*** (0.070)
Physician Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Stage Main Effect	Absorbed	Absorbed	Absorbed	Absorbed	Absorbed	Absorbed
Observations	3852	3852	1896	1896	1956	1956
Physicians	321	321	158	158	163	163

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Main sample unless stated otherwise. All models include physician and presentation effects. Low capability is a Stage 1 mean score below 7.5; high capability is at least 7.5. Both continuous moderators are standardized in the full main sample.

Supplementary materials

Beyond Augmentation: Immediate Learning from AI in Expert Decision Making

This supplement provides supporting analyses and details of the diagnostic task. Tables S1–S7 follow their first appearance in the main manuscript; the remaining tables continue consecutively as S8–S13. Section S.1 describes the cases, Section S.2 presents robustness and baseline comparisons, Section S.3 reports diagnostic score variability, and Section S.4 explains recruitment, task procedures, sample construction, and ethics.

S.1 Cases and presentation sequence

The task used 16 distinct pulmonary nodule cases with pathological diagnoses, drawn from a tertiary hospital system. Physicians reviewed 4–11 selected CT images and basic clinical information for each case, classified the nodule as AAH, AIS, MIA, or IAC, and reported diagnostic confidence. Table S1 records each presentation, its underlying case, stage and pathology, and observed performance in the main sample.

The 20 presentations comprised four Stage 1 cases (C1–C4), eight new Stage 2 cases (C5–C12), and eight Stage 3 presentations. In Stage 3, C13–C16 were new cases and C17–C20 repeated the Stage 1 cases in a different order: C17 = C2 (AAH), C18 = C1 (MIA), C19 = C4 (AIS), and C20 = C3 (IAC). Thus, presentation IDs differ across stages even when the underlying case is the same. AAH appeared in two distinct cases, C2 and C16, and was absent from Stage 2. C11 contained the known incorrect AI density label; Table S2 reports the corresponding exclusion analysis.

Table S1. Case identities and observed performance

Presentation / case	Stage / pathology	Status	No AI Exact / time	AI exposure Exact / time	AI feature note
C1 / U1	1 / MIA	New	45.5% / 45.6s	44.9% / 59.6s	—
C2 / U2	1 / AAH	New	43.3% / 35.6s	42.8% / 45.9s	—
C3 / U3	1 / IAC	New	50.0% / 23.8s	57.2% / 25.5s	—
C4 / U4	1 / AIS	New	38.1% / 26.0s	24.1% / 34.5s	—
C5 / U5	2 / MIA	New	47.0% / 40.6s	69.5% / 35.3s	—
C6 / U6	2 / AIS	New	29.9% / 32.3s	42.8% / 31.6s	—
C7 / U7	2 / AIS	New	21.6% / 27.8s	32.1% / 27.2s	—
C8 / U8	2 / IAC	New	44.8% / 27.7s	64.2% / 24.8s	—

Presentation / case	Stage / pathology	Status	No AI Exact / time	AI exposure Exact / time	AI feature note
C9 / U9	2 / AIS	New	33.6% / 28.6s	47.6% / 25.1s	—
C10 / U10	2 / MIA	New	28.4% / 28.6s	53.5% / 26.2s	—
C11 / U11	2 / MIA	New	20.9% / 27.8s	55.6% / 25.9s	Density error
C12 / U12	2 / IAC	New	56.7% / 23.2s	81.8% / 21.6s	—
C13 / U13	3 / MIA	New	45.5% / 34.6s	76.5% / 32.7s	—
C14 / U14	3 / IAC	New	63.4% / 21.5s	79.7% / 17.6s	—
C15 / U15	3 / AIS	New	27.6% / 27.1s	40.1% / 23.5s	—
C16 / U16	3 / AAH	New	27.6% / 20.3s	56.1% / 21.8s	—
C17 / U2	3 / AAH	Repeated	44.8% / 27.6s	64.2% / 22.8s	—
C18 / U1	3 / MIA	Repeated	35.1% / 19.1s	62.0% / 19.9s	—
C19 / U4	3 / AIS	Repeated	29.1% / 22.9s	43.9% / 21.9s	—
C20 / U3	3 / IAC	Repeated	68.7% / 18.5s	71.1% / 14.5s	—

Notes. Exact is the percentage of exactly correct classifications; time is mean winsorized seconds. C denotes presentation and U denotes underlying case. New cases are new at first presentation; C17–C20 repeat Stage 1 cases. N = 321.

S.2 Robustness and baseline comparisons

Tables S2–S8 present case exclusions, baseline adjustment, two-way clustering, Stage 1 outcome comparisons, alternative sample screening rules, changes within the AI-exposure cohort, and multiple-testing sensitivity. Table S2 excludes the known AI-error case from both cohorts, whereas Table S7 retains it and controls for AI_error within the AI-exposure cohort. These are separate analyses. Each table specifies its sample, outcome, and estimation method.

Table S2. Case exclusions for the known AI error and pathology coverage

Specification	Stage 2 Accuracy Score	Stage 3 Accuracy Score
Exclude C11	1.264*** (0.151)	1.479*** (0.147)
Exclude baseline AAH	1.302*** (0.177)	—

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Main sample unless stated otherwise. C11 is excluded in both cohorts' Stage 2 observations: 3,531 observations. It never enters the Stage 1–Stage 3 model, whose estimate is unchanged. The baseline AAH exclusion removes C2 from Stage 2 comparisons:

3,531 observations. All retain 321 physicians.

Table S3. Baseline adjustment and time measurement sensitivity

Model and outcome	Stage 2	Stage 3
Score DID	1.399*** (0.148)	1.479*** (0.147)
Score ANCOVA	1.216*** (0.104)	1.290*** (0.100)
Winsorized seconds DID	−10.983*** (1.887)	−10.734*** (1.837)
Winsorized seconds ANCOVA	−4.960*** (1.303)	−4.343*** (1.139)
Raw seconds DID	−25.432*** (8.007)	−24.206*** (8.282)
Log raw time DID	−0.220*** (0.045)	−0.200*** (0.044)

Notes. Standard errors: physician-clustered for DID; HC1 for ANCOVA. Stars denote unadjusted two-sided p-values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. DID uses 3,852 case observations per comparison. ANCOVA uses 321 physician observations, baseline outcome and measured characteristics listed in main-manuscript Section 6.3. Log models use raw time. ANCOVA controls the corresponding Stage 1 outcome, Gender, Age, Education categories, Professional Title categories, Nodule Diagnosis Experience, and Self-evaluation.

Table S4. Average effects with two-way clustering by physician and underlying case

Term	(1) Stage 2 Accuracy Score	(2) Stage 2 Time	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Treat × Stage	1.399*** (0.233)	−10.983*** (2.717)	1.479*** (0.290)	−10.734*** (2.486)
Two-sided p-value	<0.001	0.0019	0.0014	0.0035
Physician Fixed Effects	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	Yes	Yes	Yes	Yes
Physician clusters	321	321	321	321
Underlying-case clusters	12	12	8	8
Observations	3,852	3,852	3,852	3,852

Notes. Point estimates use the same outcomes, 321 physicians, and physician and case-presentation fixed effects as main Table 3. Time is measured in winsorized seconds. Two-way clustered standard

errors are in parentheses. Repeated presentations share their underlying-case cluster: C17 = C2, C18 = C1, C19 = C4, and C20 = C3. There are 12 underlying-case clusters for Stage 2 and 8 for Stage 3; two-sided t tests use 11 and 7 degrees of freedom, respectively. The absorbed model uses reghdfe's noconstant option, suppressing the normalization-dependent fixed-effect intercept and reporting inference for the treatment-stage slope. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table S5. Unaided Stage 1 outcomes in the main analysis sample

Stage 1 outcome	AI-exposure (n = 187)	No-AI comparison (n = 134)	Difference AI − No-AI	Welch p-value
Diagnostic Accuracy (0–10)	6.571 (1.121)	6.838 (1.184)	−0.267	0.0428
Exact-match accuracy (%)	42.246 (15.117)	44.216 (18.877)	−1.970	0.3182
Diagnostic Confidence (1–7)	5.183 (0.696)	5.032 (0.812)	0.151	0.0818
Diagnostic Time (seconds)	41.381 (22.191)	32.756 (15.956)	8.625	<0.001

Notes. N = 321 physicians. Each outcome is the physician's mean across four Stage 1 cases. Cohort entries are means (sample standard deviations). Differences equal AI-exposure minus No-AI comparison means. P-values are unadjusted, two-sided Welch unequal-variance tests using one observation per physician. Exact-match accuracy is reported as a percentage, and its difference is in percentage points. Diagnostic Time uses Stata percentile winsorization at the pooled main-sample 1st and 99th percentiles (6.03 and 166.43 seconds), consistent with the principal models.

Table S6. Average accuracy score sensitivity to sample screening

Sample	Physicians	Stage 2	Stage 3
Main sample	321	1.399*** (0.148)	1.479*** (0.147)
All completed participants	360	1.202*** (0.149)	1.296*** (0.149)
Accuracy score ≥ 5; pre-AI rapid-response screen only	351	1.209*** (0.147)	1.327*** (0.144)
Relaxed rapid screen	340	1.263*** (0.149)	1.364*** (0.145)
No score screen; all-stage rapid-response screen	325	1.391*** (0.151)	1.474*** (0.151)

Sample	Physicians	Stage 2	Stage 3
No score screen; pre-AI rapid-response screen only	356	1.203*** (0.148)	1.319*** (0.147)

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Main sample unless stated otherwise. Full retains all 360. Baseline-only requires Stage 1 accuracy score ≥ 5 and fewer than two Stage 1 times < 5 seconds. Relaxed requires Stage 1 accuracy score ≥ 5 and fewer than three rapid responses across all 20 cases. All include physician and presentation effects. The 325-person sample retains all baseline scores but applies the original all-stage time rule. The 356-person sample retains all baseline scores and excludes only physicians with at least two Stage 1 raw times below 5 seconds.

Table S7. Within-physician changes in diagnostic accuracy and time in the AI-exposure cohort

Term	(1) Stage 2 Accuracy Score	(2) Stage 2 Time	(3) Stage 3 Accuracy Score	(4) Stage 3 Time
Stage	1.030*** (0.104)	−13.970*** (1.543)	1.273*** (0.095)	−19.532*** (1.361)
AI_error	0.073 (0.205)	−1.500 (1.235)	—	—
Constant	6.571*** (0.067)	41.381*** (1.032)	6.571*** (0.063)	41.381*** (0.907)
Physician Fixed Effects	Yes	Yes	Yes	Yes
Case-presentation Fixed Effects	No	No	No	No
Observations	2,244	2,244	2,244	2,244
Physicians	187	187	187	187

Notes. Each column uses four Stage 1 cases and eight target-stage cases for each of 187 AI-exposure physicians (2,244 observations). Stage equals zero in Stage 1 and one in the target stage. All models include physician fixed effects; no case-presentation fixed effects or Self-evaluation interactions are included. Treat is constant at one and is omitted. AI_error equals one for the known incorrect AI density label in C11 and zero otherwise. C11 is retained in columns (1)–(2); in columns (3)–(4), AI_error has no variation and is omitted (—). Diagnostic Accuracy is scored 0, 5, or 10. Diagnostic Time is winsorized at the pooled main-sample 1st and 99th percentiles and measured in seconds. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. These coefficients describe changes within the AI-exposure cohort and do not subtract changes in the No-AI comparison cohort.

Table S8. Multiplicity sensitivity of self evaluation interactions

Stage	Interaction (SE)	Unadjusted p	Holm p 2 tests	Holm p 4 tests
Stage 2	−0.287** (0.133)	0.0314	0.0314	0.0942
Stage 3	−0.338** (0.132)	0.0108	0.0217	0.0434

Notes. Two-test family contains the two self-evaluation triple interactions. Four-test family adds the capability triple interactions in the capability-only models.

S.3 Diagnostic score variability

Table S9 describes within-physician SDs and ranges and between-physician dispersion. Table S10 estimates changes in within-physician variability using 642 physician-stage observations per comparison. Table S11 compares the same four underlying cases across Stages 1 and 3, holding case identity and case count fixed. These dispersion models include physician fixed effects and a separate target-stage dummy.

Table S9. Within and between physician accuracy score dispersion

Stage	Cohort	Mean within SD	Mean within range	Between SD of means	Min–max of means
1	No AI	3.097	6.119	1.184	5.000–10.000
1	AI exposure	3.404	6.818	1.121	5.000–8.750
2	No AI	2.777	6.642	0.962	5.000–8.750
2	AI exposure	2.768	6.043	1.009	3.750–10.000
3	No AI	3.229	8.060	0.981	5.000–8.750
3	AI exposure	2.900	6.551	0.850	5.000–10.000

Notes. Within SD and range are calculated across cases for each physician, then averaged by cohort-stage. Between SD is the SD of physicians' stage mean scores within a cohort. Stage 1 has four cases; each later stage has eight.

Table S10. Cohort changes in within physician dispersion

Term	(1) Stage 2 SD	(2) Stage 2 Range	(3) Stage 3 SD	(4) Stage 3 Range
Treat × Stage	−0.316** (0.146)	−1.298*** (0.378)	−0.635*** (0.141)	−2.208*** (0.372)
Observations	642	642	642	642
Physicians	321	321	321	321

Notes. Physician-clustered standard errors are in parentheses. Stars denote unadjusted two-sided p-

values: * p < 0.10, ** p < 0.05, *** p < 0.01. Main sample unless stated otherwise. Unit is physician-stage. Each model includes physician effects and an independent target-stage dummy. Outcomes are score points, not standard errors.

Table S11. Dispersion changes using the same four underlying cases

Outcome	Treat × Stage 3 (SE)	95% CI	p-value	Observations
Within-physician SD	−0.629*** (0.174)	[−0.971, −0.287]	<0.001	642
Within-physician range	−1.471*** (0.394)	[−2.245, −0.697]	<0.001	642

Notes. Only Stage 1 C1–C4 and Stage 3 C17–C20 are used, with four scores per physician at each stage. C17 = C2, C18 = C1, C19 = C4, and C20 = C3. Models include physician effects and a separate Stage 3 dummy, with physician-clustered standard errors; N = 321 physicians. This keeps the case set and case count fixed. It does not eliminate repetition, bounded-score effects or nonrandom cohort confounding. Stars are unadjusted: * p < 0.10, ** p < 0.05, *** p < 0.01.

S.4 Experiment implementation and descriptive information

S.4.1 Recruitment and baseline questionnaire

We recruited licensed physicians in medical imaging departments or related specialties at tertiary hospitals through a professional online platform. Participants worked in 153 Chinese cities. Eligibility required experience interpreting chest CT scans for pulmonary nodules and familiarity with the four diagnostic categories. The platform verified professional credentials, hospital affiliation, and specialty. All batches used the same eligibility criteria, task materials, and fixed compensation of RMB 370, which did not depend on diagnostic performance.

Over a seven-month period in 2025, we randomly recruited 200 eligible physicians from tertiary hospitals in China for the AI-exposure cohort, followed by 160 eligible physicians for the No-AI comparison cohort. All physicians completed the same presentations in the same order within one session. The initial sample contained 360 physicians.

Before the diagnostic task, physicians reported gender, age, city, specialty, education, professional title, years of pulmonary nodule diagnosis experience and self-evaluation relative to peers. The questionnaire reproduced below shows the response categories and scale anchors.

S.4.2 Task sequence and information shown to physicians

In Stage 1, all physicians diagnosed C1–C4 without AI. In Stage 2, physicians diagnosed eight new cases, C5–C12; only the AI-exposure cohort received AI-assisted information. In Stage 3, AI assistance was removed and both cohorts diagnosed C13–C20 without AI. Four Stage 3 cases were new and four repeated Stage 1 cases, as documented in Table S1. Short breaks preceded Stages 2 and 3. Table S12 reports the stage means by cohort.

The common interface displayed selected CT images and patient age, gender, smoking history, and relevant medical history. Physicians could scroll through the images and use Ctrl plus the mouse wheel to zoom. The AI-assisted interface additionally showed nodule location, size, density, pleural indentation, vessel convergence, and other imaging features. AI did not supply an AAH/AIS/MIA/IAC classification, a category probability, or a recommended final diagnosis. Physicians made the final classification themselves. Neither cohort received the pathological diagnosis or feedback about whether its answers were correct. Figure S1 illustrates the information available to physicians under the No-AI and AI-assisted conditions, using Case 5 as an example.



Figure S1. Diagnostic Task Interfaces With and Without AI Assistance

Notes. Both panels show Case 5. The left panel displays the CT image and clinical information presented without AI assistance. The right panel additionally displays a nodule localization marker and AI-provided information on nodule location, size, density, and imaging signs.

Diagnostic Accuracy was scored as 10 for an exact match to pathological truth, 5 for a diagnosis one category away in the AAH–AIS–MIA–IAC ordering, and 0 otherwise. Exact-match accuracy was coded as one for an exactly correct diagnosis and zero otherwise. After classifying each case, physicians reported diagnostic confidence on a 1–7 scale. Diagnostic Time measured the seconds from presentation

of the case page to completion of the classification and confidence questions. The tables report time winsorized at the pooled main-sample 1st and 99th percentiles where specified; the rapid-response screening rule used raw times.

S.4.3 Sample construction and ethics

The main-sample rules excluded physicians whose Stage 1 mean diagnostic score was below 5 or who had at least two raw case times below five seconds across all 20 presentations. Six physicians met the score criterion and 35 met the time criterion; two met both. Taking the union excluded 39 physicians and left 321: 187 in the AI-exposure cohort and 134 in the No-AI comparison cohort. Table S13 reports the flow by cohort. Table S6 presents alternative screening rules; the additional baseline exact-match threshold in main Table 5 is a separate sensitivity analysis.

This study was conducted in compliance with all relevant ethical regulations and received approval from the University of Toronto Social Sciences, Humanities and Education Research Ethics Board (protocol 67817; RIS 51226). Patient identifiers in the case materials were encrypted. The study was not a clinical trial and was not publicly preregistered.

Table S12. Stage means by cohort

Stage	Cohort	N	Accuracy Score	Exact %	Confidence	Time seconds
1	No AI	134	6.838	44.216	5.032	32.756
1	AI exposure	187	6.571	42.246	5.183	41.381
2	No AI	134	6.479	35.354	5.035	29.582
2	AI exposure	187	7.610	55.882	5.473	27.223
3	No AI	134	6.632	42.724	5.054	23.958
3	AI exposure	187	7.844	61.698	5.368	21.849

Notes. Means of physician-stage averages. Stage 1 contains four cases; Stages 2 and 3 each contain eight. Time is winsorized.

Table S13. Recruitment and sample construction

Cohort	Initial	Accuracy Score < 5	Fast ≥ 2	Overlap	Main
AI exposure	200	5	10	2	187
No AI	160	1	25	0	134

Cohort	Initial	Accuracy Score < 5	Fast ≥ 2	Overlap	Main
Total	360	6	35	2	321

Notes. Fast counts raw case times below five seconds across 20 cases. Exclusion is the union of the two flags.

S.4.4 Questionnaire and task instructions

(1) Pre-experiment Questionnaire

Q1 Please select your gender. [Single choice]

- ○ Male
- ○ Female

Q2 Please select your highest educational degree. [Single choice]

- ○ Associate degree
- ○ Bachelor's degree
- ○ Current undergraduate student
- ○ Master's degree
- ○ Current master's student
- ○ Doctoral degree
- ○ Current doctoral student

Q3 Please select your professional title. [Single choice]

- ○ Junior
- ○ Intermediate
- ○ Associate senior
- ○ Senior
- ○ None
- ○ Other

Q4 Please enter your age in years. [Open-ended]

Q5 Please enter the city where you work. [Open-ended]

Q6 Please enter your specialty. [Open-ended]

Q7 How many years of experience do you have in lung nodule diagnosis? Unit: years. [Open-ended]

Q8 How long have you used AI assistance for lung nodule diagnosis? Unit: months. [Open-ended]

Q9 How confident are you in completing lung nodule diagnostic tasks? [Scale]

[Not confident at all, Very confident]

1 2 3 4 5 6 7

Q10 Compared with your peers, how would you rate your professional competence in lung nodule diagnosis? [Scale]

[Far below peers, Far above peers]

1 2 3 4 5 6 7

（2）Questions During the Experiment

① stage1

First, physicians were informed that they needed to complete diagnostic tasks for four cases. For each case, only several key CT images were selected, ranging from 4 to 11 images, and each case contained only one nodule. On the computer, participants could place the cursor over the image, scroll through the CT images using the mouse wheel, and zoom in using Ctrl + mouse wheel.

The case images and clinical information were then displayed, including patient age, gender, smoking history, and relevant medical history.

Physicians then answered the following questions:

Q1 Please diagnose which of the following types is most likely for the patient's lung nodule. [Single choice]

- ○ Atypical adenomatous hyperplasia (AAH)
- ○ Adenocarcinoma in situ (AIS)
- ○ Minimally invasive adenocarcinoma (MIA)
- ○ Invasive adenocarcinoma (IAC)

Q2 In this case, how accurate do you think your diagnosis is? [Scale]

[Completely inaccurate, Completely accurate]

1 2 3 4 5 6 7

② stage2

Physicians were first asked to take a short break and then complete lung nodule type diagnoses for eight cases.

For physicians in the non-AI-assisted group, the procedure was the same as in Stage 1.

For physicians in the AI-assisted group, the case images and clinical information were displayed, together with AI-assisted information about lung nodules, including their location, size, density, and other imaging features.

Physicians then answered the following questions:

Q1 Please diagnose which of the following types is most likely for the patient's lung nodule. [Single choice]

- ○ Atypical adenomatous hyperplasia (AAH)
- ○ Adenocarcinoma in situ (AIS)
- ○ Minimally invasive adenocarcinoma (MIA)
- ○ Invasive adenocarcinoma (IAC)

Q2 In this case, how accurate do you think your diagnosis is? [Scale]

[Completely inaccurate, Completely accurate]

1 2 3 4 5 6 7

③ stage3

Physicians were first asked to take a short break and then complete lung nodule type diagnoses for eight cases.

The procedure was the same as in Stage 1.