Blueprint Labs

Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology

Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz

Policy Brief | September 2023

Summary

In recent years, artificial intelligence (AI) in medicine has focused on inventing and refining algorithms as a diagnostic tool. In radiology, AI has already matched or surpassed the accuracy of humans. Geoffrey Hinton, the godfather of AI and a Turing Award winner, famously suggested in 2016 that deep learning will supplant radiologists. Other experts, though, predict that radiologists will more likely collaborate with AI than be replaced by it.

Nikhil Agarwal (MIT), **Alex Moehring** (MIT), **Pranav Rajpurkar** (Harvard), and **Tobias Salz** (MIT) experimented with various frameworks for using AI in radiology, with and without human collaboration. They randomly assigned AI support to radiologists to explore how they use AI predictions in their diagnoses.

They found that on its own, AI was more accurate in its predictions than nearly two-thirds of radiologists. On average, even when offered access to AI, radiologists' accuracy did not improve. Looking only at the average impact of AI on radiologists does not tell the full story, however. Not all AI predictions have the same effect on radiologists – the AI's confidence matters. If the AI predicts a certain pathology as very likely (>80%) or unlikely (<20%), it is considered "confident." Confident AI predictions improve radiologists' accuracy, while uncertain AI reduces it. Radiologists' confidence also matters: AI assistance is harmful when radiologists are certain that a pathology is very unlikely.

The disparate impacts of AI predictions can be explained by two types of mistakes that radiologists make. First, radiologists underweight AI predictions compared with their own baseline evaluations. Second, AI and radiologist predictions are somewhat correlated, but radiologists don't take this correlation into account. Moreover, radiologists take longer to make decisions when they receive AI assistance.

Given these findings, how can AI-radiologist collaboration be designed to generate the most accurate diagnoses? The researchers find that, depending on the confidence of the AI prediction, cases should be assigned *either* to AI or to radiologists, because uncertain AI predictions lead radiologists astray. In other words, in most cases AI-radiologist collaboration is ineffective.

These findings show that to fully exploit the promises of AI, researchers and policymakers need to better understand how humans use and misuse it. This work offers critical lessons on how to design systems between AI and humans in radiology and beyond.

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023): "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," *MIT Blueprint Labs Discussion Paper #2023.10*.

Background and Policy Relevance

AI has the potential to displace humans from tasks that require complex reasoning. However, many experts suggest that because of ethical and legal concerns radiologists are more likely to collaborate with AI than to be replaced by it.

How radiologists use AI predictions, whether AI improves their performance, and how to design AI-radiologist collaboration are all open questions. Radiology offers a particularly useful laboratory to seek answers. The radiologist experimental setting is nearly identical to how they work day-to-day. It is an example of a highly skilled profession that is being transformed by AI. Further, since radiologists are highly paid, improved productivity could lead to large savings.

Setting and Methods

The researchers conducted a remote experiment with radiologists who analyzed chest X-rays. They recruited 180 professional radiologists through teleradiology companies to diagnose retrospective patient cases. They designed an experiment to compare how radiologists perform when given access to different levels of information. It consisted of four treatment groups: Besides the chest X-ray, one group received AI predictions; a second, clinical histories; a third, both AI and clinical histories; and the fourth, no additional information.

To evaluate radiologist-AI collaboration, the researchers carried out randomized control trials that varied the order of treatments and the number of times each radiologist reviewed the same X-ray: this allowed for comparison of radiologists' accuracy under different settings. The experiment data contains 324 historical cases. The AI-predictions were computed using CheXpert, an AI algorithm trained with 224,316 cases of 65,240 patients labeled for the presence of common chest pathologies. The AI provides the probability that a given chest pathology is positive (from 0-1). The radiologists provided the probability of a given chest pathology and a recommendation of whether to treat or follow up.

For privacy reasons, AI was not trained using clinical histories, which typically includes patients' vitals and labs.

One challenge is that definitive diagnostic tests do not exist for most chest pathologies. To evaluate the quality of the diagnoses, the researchers construct a diagnostic standard using the majority prognosis from a group of five board-certified radiologists. Radiologists' accuracy is measured against this benchmark.

Key finding #1: Though AI is more accurate than the majority of radiologists, AI assistance does not, on average, improve radiologists' diagnostic accuracy.

On their own, AI predictions were more accurate than those of nearly two-thirds of radiologists. If humans correctly incorporated AI predictions with their own information, then AI assistance would unambiguously improve their accuracy. The study, though, finds that access to AI on average did *not* improve radiologists' performance.

It is not the case that radiologists ignore AI – in fact, in most cases, their predictions move toward AI's. Rather, the overall findings mask important heterogeneity: effects of AI assistance on diagnostic accuracy depends on the interrelation between confidence levels of AI and humans.

If the AI predicts a certain pathology is very likely (>80%) or very unlikely (<20%), it is

considered "confident." When AI is confident that a pathology is very *unlikely*, it helps radiologist accuracy. But when AI is confident that a pathology is very *likely*, it has no discernible impact on radiologist accuracy. When AI is uncertain, it tends to hurt radiologist performance (see Figure 1a).

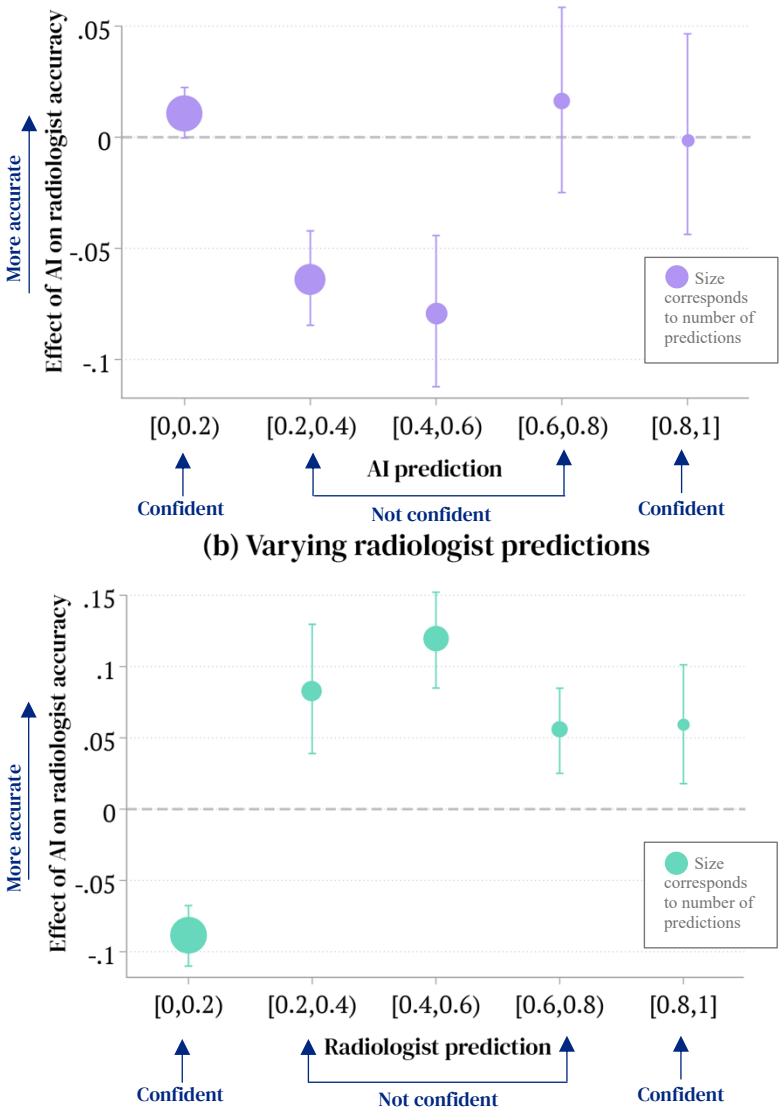
Similarly, the confidence level of radiologists matters. AI support helps unconfident radiologists and those who predict a pathology is very likely. AI hurts the accuracy of radiologists who predict the pathology is very unlikely (see Figure 1b). Notably, the vast majority of cases fall into this final category, which explains why AI has a negative average impact on radiologist accuracy. Finally, when clinical history is provided, radiologist accuracy improves.

Key finding #2: Radiologists do not correctly combine their information with AI predictions, diminishing the potential benefit of access to AI.

Two types of human error help explain this finding. First, radiologists underweight AI assistance relative to their baseline evaluation, a phenomenon referred to as "automation neglect."

Second, radiologists are susceptible to treating their baseline predictions as independent of AI predictions even though this is not the case. Indeed, the predictions are highly correlated since they are both based on the same X-ray. This misjudgment, known as "correlation neglect," may lead radiologists to be overconfident when the AI prediction agrees with their own. These two behavioral biases greatly diminish the potential benefits of AI assistance.

Figure 1: Effect of AI on Radiologists' Accuracy



How to read this figure: On the x-axis in 1(a), cases are categorized into one of five groups, based on the confidence of the AI prediction (e.g., if the AI predicts a 0.1 chance of a given pathology, it is placed in the first group). The first and last groups (close to 0 and 1) are considered "confident." Points on the y-axis refer to the impact of the AI on the radiologist's accuracy (more positive means more accurate). Accuracy is measured as the negative of the distance from the diagnostic standard – further from the diagnostic standard is less accurate. When vertical bars straddle zero, AI assistance has no observable effect on a radiologist's accuracy. In Figure 1(b), the groups on the x-axis are constructed using the radiologists' predictions.

Key finding #3: To maximize accuracy, patient cases should be delegated to either AI or radiologists, but an AI assisted radiologist is suboptimal.

The researchers evaluated various forms of human-AI collaboration. Radiologists using AI take 4% more time per case, making decisions less efficient. This additional time, coupled with the aforementioned behavioral biases, argues against having radiologists make decisions *with* AI assistance.

Here are two scenarios for deciding whether to bring in a radiologist after the AI algorithm makes an initial prediction based on a chest X-ray.

- If the AI prediction is very confident (e.g., <20% predicted probability of a pathology), the diagnosis would be relied upon without radiologist review.
- If the AI prediction is less confident, the chest X-ray would be reviewed by a radiologist *without* access to the AI prediction. If the radiologist's diagnosis agrees with the AI's, then that diagnosis would be used. If the diagnoses differ, the case would be escalated for a second radiologist's opinion.

Under these scenarios, AI could often provide a prediction as soon as the diagnostic image is captured, with only a subset of cases calling for a radiologist's input. This approach would allow for more timely diagnoses and reduce overall costs.

Future Research

As AI continues to reshape the nature of work across a range of fields, it is critical to continue to uncover the benefits and pitfalls of human-machine collaboration. Within radiology, the human biases discussed above will need to be addressed to exploit the potential benefits of AI-human collaboration.

The researchers plan to study whether radiologists who have more experience using AI perform better and whether AI-specific training for radiologists leads to improved outcomes.

They also hope to identify the type of AI predictions that are especially likely to be incorrect and if radiologists could reliably correct such AI diagnoses. If they cannot, then preemptively withholding these predictions may be preferable.

Finally, the researchers plan to monitor the impacts of the evolving regulatory landscape.