Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology
Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz
Policy Brief | August 2023

Summary

In recent years, artificial intelligence (AI) in medicine has focused on inventing and refining algorithms that accomplish complex diagnostic tasks with high accuracy. Radiology is an important specialty where such algorithms are already applied and where AI tools have matched or surpassed the performance of human radiologists. Geoffrey Hinton, the Godfather of AI and Turing Award winner, famously suggested in 2016 that, thanks to deep learning, we will no longer need to train radiologists. Yet, other experts propose that radiologists are more likely to collaborate with, rather than be replaced by, AI. Ethical, legal, and regulatory challenges may also hinder complete AI automation.

Researchers Nikhil Agarwal (MIT), Alex Moehring (MIT), Pranav Rajpurkar (Harvard), and Tobias Salz (MIT) conducted an experiment to examine the ideal design of radiologist-AI collaboration and its impacts. They randomly assigned AI support to radiologists to explore how they use AI predictions in their diagnoses.

On their own, AI predictions were more accurate than nearly two-thirds of radiologists. However, when radiologists were given access to AI, their average performance did not improve.

Looking only at the average impact of AI on radiologists masks important differences. Not all AI predictions have the same effect on radiologists – the AI’s confidence matters. If AI predicts a certain pathology is extremely likely (>80%) or unlikely (<20%), it is considered “confident.” Confident AI predictions improve radiologists’ accuracy, while uncertain AI hurts radiologists’ performance.

The disparate impacts of AI predictions can be explained by two types of mistakes that radiologists make. First, radiologists fail to accurately weigh AI predictions, exhibiting “automation neglect.” They underweight AI predictions compared with their own baseline evaluations. Second, radiologists treat AI predictions and their own evaluations as independent, even though they are not. Moreover, radiologists take longer to make decisions when they receive AI assistance. These behavioral biases and the increased time diminish the potential benefits of AI-radiologist collaboration.

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

These findings provide valuable lessons for researchers and policymakers eager to integrate AI into healthcare. While AI holds promise, it is critical to better understand how humans use it, the errors they make, and the opportunities that lie ahead. This work offers critical lessons on how to design systems between AI and humans in radiology and beyond.

**Source**

**Background and Policy Relevance**

AI has the potential to displace humans from tasks that require complex reasoning. In recent years, some AI tools have demonstrated the ability to match or surpass the accuracy of human radiologists. However, many experts suggest that radiologists are more likely to collaborate with, rather than be replaced by, AI due to ethical, legal, and regulatory challenges. Amid these challenges, radiologist-AI collaboration has the potential to offer clinical gains that cannot be achieved by radiologists or AI alone.

How radiologists use AI predictions, whether AI improves their performance, and how to design AI-radiologist collaboration are all open questions. Radiology offers a valuable laboratory to examine these questions: It is an iconic example of a highly-skilled profession that is being transformed by AI. In addition, radiology is highly paid, so potential advancements in productivity could have large financial implications.

Economics offers a powerful tool to shed light on how humans interact with AI. This knowledge can help shape the institutions that guide the use of AI to ensure that its development is beneficial to society.

**Setting and Methods**

The researchers conduct a remote experiment with radiologists who analyze chest X-rays. They recruited 180 professional radiologists through teleradiology companies to diagnose retrospective patient cases. The experiment consists of four treatment groups: each group receives a chest X-ray with AI predictions, clinical history information, both, or neither. Radiologists typically have access to clinical history information, such as the patient’s vitals and labs. However, for privacy reasons, the AI was not trained using this clinical information.

To evaluate radiologist-AI collaboration, the researchers employ three experimental designs that vary the order of treatments and the number of times each radiologist reviews the same X-ray. The experiment data contains 324 historical cases. The AI provides the probability that a given chest pathology is positive (from 0-1).

The researchers then analyze the radiologists’ diagnoses in each treatment group: Radiologists provide the probability of a given chest pathology and a recommendation of whether to treat or follow up.

---

1 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.
One challenge is that definitive diagnostic tests do not exist for most chest pathologies. To evaluate the quality of the diagnoses, the researchers construct ground-truth labels using the majority prognosis from a group of five board-certified radiologists. Radiologists’ accuracy is measured by comparing to these ground-truth labels.

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

On their own, AI predictions were more accurate than nearly two-thirds of radiologists. If humans correctly incorporated AI predictions with their own information, then AI assistance would unambiguously improve their accuracy. However, when radiologists were given access to AI, their average performance did not improve.

This lack of effect is not because radiologists ignore AI – their predictions move toward AI’s. Rather, the overall effect masks important heterogeneity. In other words, not all AI predictions have the same effect on radiologists.

If the AI predicts a certain pathology is very likely (close to one) or very unlikely (close to zero), it is considered “confident.” Confident AI predictions improve radiologists’ accuracy, while uncertain AI hurts radiologist performance (see Figure 1a).

Similarly, radiologists’ certainty matters. AI support helps unconfident radiologists but hurts the accuracy of confident ones (see Figure 1b). In contrast, providing clinical history improves all radiologists’ accuracy, suggesting that humans can currently access valuable information that AI assistance cannot.

---

2 For AI predictions close to one, there is not a significant effect on radiologists’ accuracy. This is because there are fewer positive cases (prediction close to one) in the dataset.
Key finding #2: Radiologists do not correctly combine their information with AI predictions, diminishing the potential benefit of access to AI.

AI predictions perform better on their own than two-thirds of radiologists. Yet, radiologists do not improve when provided with AI assistance. Two types of human error help explain this finding.

First, radiologists do not weigh their own information correctly, as compared with the AI's. More specifically, they underweight AI assistance by approximately 30% relative to their baseline evaluation (a phenomenon referred to as “automation neglect”).

Second, AI predictions are highly correlated with radiologists’ baseline predictions. Yet, radiologists ignore this correlation. In other words, they act as if their own information and AI predictions are independent, even though they are not (referred to as “correlation neglect”). These two behavioral biases greatly diminish the potential benefits of AI assistance.

Key finding #3: To maximize accuracy and minimize physician effort, patient cases should be delegated to either AI or radiologists, but not both together.

The researchers evaluate various forms of human-AI collaboration. Radiologists using AI take 4% more time per case, making decisions less efficient. This additional time, coupled with radiologists’ behavioral biases, work against having radiologists make decisions with AI assistance. The researchers find that cases should either be assigned to the radiologist or to AI, but rarely to both together. In other words, from a diagnostic performance perspective, radiologists should work next to rather than with AI.

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 for the potential benefits of AI-human collaboration to be realized.

Future Research

To further investigate the potential gains of AI assistance, the researchers plan to study whether radiologists who have more experience using AI perform better. Similarly, they hope to examine whether AI-specific training for radiologists leads to improved outcomes. They also aim to explore whether they can identify AI predictions that are especially likely to be incorrect. They can then analyze whether radiologists can correct these incorrect AI diagnoses – if not, preemptively withholding these predictions could be valuable. Finally, examining the impacts of the evolving regulatory landscape is a pressing topic for future work.